Raffaello Cortina Editore

Mariarosaria Taddeo

Codice di guerra

Etica dell'intelligenza artificiale nella difesa

Scienza e idee

Collana fondata da Giulio Giorello

La digitalizzazione della difesa nazionale, iniziata con la prima guerra del Golfo, ha trovato nella guerra in Ucraina il suo punto di non ritorno. In questi anni è diventato evidente che la grande quantità di dati che produciamo ogni giorno unita all'intelligenza artificiale (IA) ha una notevole importanza nei processi decisionali e operativi della difesa e sicurezza nazionale: dall'intelligence alle operazioni cibernetiche a quelle cinetiche (i veri e propri combattimenti). Tuttavia, al potenziale dell'IA si accompagnano seri rischi etici, sociali e legali, che spaziano dalle difficoltà di attribuire la responsabilità per le azioni compiute dai sistemi IA alla loro limitata predicibilità e sicurezza, al problematico rapporto tra uso dell'IA nella difesa e Teoria della Guerra Giusta. Se la difesa è il banco di prova del rispetto dei valori democratico-liberali, l'adozione dell'IA in questo settore non può prescindere da una valutazione di tipo etico per un suo sviluppo responsabile.

Mariarosaria Taddeo presenta un'analisi concettuale e sistematica dei problemi che derivano dall'uso dell'IA, aprendo il dibattito su opportunità e rischi per la difesa e offrendo raccomandazioni pratiche a decisori politici e professionisti. Perché senza una governance etica dell'IA sarà impossibile mitigarne la pericolosità.

Mariarosaria Taddeo è docente di Digital Ethics and Defence Technologies all'Oxford Internet Institute dell'Università di Oxford. I suoi studi vertono sull'etica e la governance delle tecnologie digitali, in particolare nel campo della sicurezza nazionale e della difesa.

Scienza e idee

Collana fondata da Giulio Giorello

Dal catalogo

Luciano Floridi

La quarta rivoluzione

Come l’infosfera sta trasformando il mondo

Luciano Floridi

Pensare l’infosfera

La filosofia come design concettuale

Luciano Floridi

Etica dell’intelligenza artificiale

Sviluppi, opportunità, sfide

Luciano Floridi

Filosofia dell’informazione

Giovanni Ziccardi

Internet, controllo e libertà

Trasparenza, sorveglianza e segreto nell’era tecnologica

Giovanni Ziccardi

Dati avvelenati

Truffe, virus informatici e falso online

Mariarosaria Taddeo

Codice di guerra

Etica dell’intelligenza artificiale nella difesa

Raffaello Cortina Editore

www.raffaellocortina.it

Titolo originale
The Ethics of Artificial Intelligence in Defence

Traduzione
Virginio B. Sala

Copertina
mara scanavino project | Alberto Lameri

ISBN 978-88-3285-829-7

Milano, via Rossini 4

Prima edizione: 2025

INDICE

PREFAZIONE

> Rispondigli che si sbaglia. La guerra non cancella il rispetto. Anzi, in guerra è ancora più necessario che in tempo di pace.
>
> KHALED HOSSEINI,
> *Il cacciatore di aquiloni*

Le basi di molte delle idee presentate in questo libro sono state poste più di quindici anni fa, quando ho avviato un progetto di ricerca sull'etica della guerra dell'informazione.[1] Il progetto del libro, però, è emerso solo più di recente, influenzato dal mio coinvolgimento in discussioni con varie organizzazioni della difesa sui temi etici relativi all'intelligenza artificiale (IA). Nel corso di queste collaborazioni, ho visto crescere l'esigenza di un quadro di riferimento che permettesse di identificare, analizzare e affrontare in modo sistematico e coerente le sfide etiche che sorgono dai molti usi correnti (ed emergenti) dell'IA nel settore della difesa, e che fungesse anche da guida per la loro governance. Con questo libro cerco di soddisfare quell'esigenza.

Clausewitz ha scritto: "La guerra è […] un atto di forza che ha per iscopo di costringere l'avversario a sottomettersi alla nostra volontà. La forza si arma delle invenzioni delle arti e delle scienze" (1832, p. 19). Sono convinta che questa sia una concezione troppo semplicistica del ruolo che hanno le tecnologie nella conduzione di una guerra. Queste non sono semplici strumenti per ottimizzare l'uso della forza; sono anche fattori dirompenti nella nostra concettualizzazione della guerra e delle sue implicazioni etiche e legali. Prendiamo, per esempio, la tecnologia nucleare e le armi nucleari. Come armi di distruzione di massa, hanno annullato la distinzione fra combattenti e non combattenti, e così facendo hanno ribaltato un'idea della guerra in cui il principio di discriminazione era fondamentale e sul quale fino a quel momento si era basata la Teoria della Guerra Giusta. Analogamente, l'uso dei droni ha contribuito a una

ridefinizione della nostra idea della guerra come “duello su vasta scala” (*ibidem*), superando l’idea di un confronto simmetrico in cui entrambe le parti corrono rischi simili (Steinhoff, 2013; Braun, Brunstetter, 2013; Strawser, 2013; Schulzke, 2016). Per questo sono d’accordo con Clark che, “quando applichiamo l’etica alla guerra, siamo costretti a puntare a un bersaglio in costante movimento” (2015, p. 19).

L’adozione di tecnologie digitali, in particolar modo dell’IA, nella difesa ha spostato il bersaglio ancora una volta ridefinendo la nostra comprensione di concetti rilevanti (Floridi, 2014). Consideriamo, per esempio, la nostra concezione della guerra come comportamenti coercitivi veicolati dall’uso della forza. Per secoli, dai tempi di Cicerone (106-43 a.C.), abbiamo regolato la guerra regolando l’uso della forza (Cicero, 2008). Dal 2014 (quando la NATO[2] ha dichiarato il cyberspazio un ambito della guerra) questo approccio normativo non funziona per tutti i casi di conflitto, perché la guerra cibernetica separa la coercizione dall’uso della forza. Purtroppo, la guerra cinetica si combatte ancora, ma negli ultimi decenni la guerra cibernetica è andata diffondendosi, e sono sorti problemi urgenti su come regolamentarla. Urgono, per esempio, regole o misure per limitare i rischi di escalation dei conflitti cibernetici, e regole per limitare le violazioni dei diritti individuali e le minacce alle infrastrutture civili.

La ridefinizione del concetto di guerra continua con l’adozione dell’IA nella difesa. Pensiamo alla famosa trinità di Clausewitz, che descrive la guerra come composta dalla “*violenza* originale”; dal “giuoco delle probabilità e del caso, che le imprimono il carattere di una *libera attività dell’anima*”; dalla “sua natura subordinata di strumento politico, ciò che la riconduce alla *pura e semplice ragione*” (1832, p. 40, corsivo mio). L’uso dell’IA ha un impatto sull’elemento violento della trinità, poiché può eliminare la violenza (forza) dalla guerra, per esempio migliorando l’efficacia delle operazioni cibernetiche non cinetiche e rendendo meno necessarie quelle cinetiche. Al contempo, l’IA può modificare il modo in cui si esercita la violenza, eliminando gli agenti umani dal processo di applicazione della forza, se/quando vengono utilizzati sistemi d’arma autonomi. L’IA può supportare il ragionamento umano migliorando la *situational awareness* (la consapevolezza situazionale, del contesto) e sostenendo il processo decisionale. Così facendo, può de-antropomorfizzare la conduzione della guerra e ostacolare la “libera

attività" a cui fa riferimento Clausewitz, se gli agenti umani accettano acriticamente le raccomandazioni di un sistema IA.

In questo libro considero l'uso dell'IA nella difesa come il caso più recente e significativo della digitalizzazione della difesa. L'IA sta ridefinendo non solo il modo in cui vengono condotte le guerre, ma anche il funzionamento delle organizzazioni della difesa, i processi decisionali e operativi, come i modi di acquisizione di dati e informazioni e le tattiche e strategie di guerra. Qui, considero *tutto* lo spettro dei cambiamenti, delle loro implicazioni concettuali ed etiche. Per questo il libro è sull'etica dell'IA nella difesa, e non sull'etica dell'IA nella guerra.

Non mi pronuncio in merito alla questione se si debba considerare l'uso delle tecnologie digitali e dell'IA nella difesa come una rivoluzione negli affari militari – limitata ad aspetti prevalentemente operativi – o come la scintilla di una rivoluzione militare. Questa è l'idea che i cambiamenti negli affari militari, nelle strategie, nelle tattiche e nella condotta della guerra provocati dall'adozione delle tecnologie digitali avranno come risultato uno spostamento nell'equilibrio delle forze e nel modo in cui vengono risolti i conflitti. La mia attenzione è invece tutta sulle implicazioni concettuali ed etiche di questa digitalizzazione e su come gestire i rischi etici e le opportunità che ne derivano; non importa poi se la digitalizzazione della difesa rientri in una categoria, nell'altra o in nessuna delle due.

La validità di questo libro nel tempo dipenderà dalla sua capacità di approfondire gli aspetti concettuali e normativi di un fenomeno contemporaneo, ossia la rivoluzione digitale, e di identificare le sue ramificazioni per il futuro prevedibile, evitando al contempo di preoccuparsi eccessivamente della cronaca di eventi specifici. Per questa ragione, l'analisi si concentra su aspetti concettuali ed etici della trasformazione digitale della difesa e dell'uso dell'IA in questo dominio ed evita di fare ricorso a resoconti descrittivi o aneddotici.

Il mio obiettivo in questo libro è presentare un quadro di riferimento etico per identificare, analizzare e affrontare sistematicamente le implicazioni etiche dei diversi usi dell'IA nel contesto della difesa, in modo da coadiuvare e contribuire a guidarne la governance. Per raggiungere questo obiettivo, mi baso sull'autonomia e capacità di apprendimento dell'IA e sul livello di agency che queste determinano, sull'etica dell'IA e sulla Teoria della Guerra Giusta.

Un contributo fondamentale di questo libro è la traduzione dell'analisi concettuale e normativa in raccomandazioni per la governance dell'IA nella difesa. In tal senso, il libro rispecchia l'idea che l'etica dell'IA funzioni al meglio come *etica translational*, capace cioè di "tradurre" analisi concettuali in linee guida per chi deve regolare la progettazione, lo sviluppo e l'uso delle tecnologie digitali (Taddeo, Floridi, 2018b, p. 752). L'approccio *translational* richiede un'analisi il più chiara possibile, del resto una traduzione oscura dal concettuale al pragmatico avrebbe poco valore. Per questo ho fatto ogni sforzo per conservare semplicità sia nel linguaggio sia nell'analisi. Ciononostante, devo ammettere che questo non è un testo introduttivo sull'etica della difesa, sull'etica militare o l'etica dell'IA, bensì un testo rivolto a un pubblico con un certo grado di competenza in etica dell'IA e Teoria della Guerra Giusta.

L'analisi si sviluppa seguendo tre categorie di usi dell'IA nella difesa (Taddeo et al., 2021): di sostegno e supporto; conflittuali e non cinetici; conflittuali e cinetici. La categoria "di sostegno e supporto" si riferisce all'impiego dell'IA per sostenere le funzioni di back-office, la logistica e quegli usi che mirano a migliorare la sicurezza delle infrastrutture e dei sistemi di comunicazione da cui dipendono i servizi di difesa nazionale. Questa categoria include anche l'uso dell'IA a sostegno di processi decisionali strategici e di pratiche di *wargaming* (ossia le simulazioni belliche). Gli usi "conflittuali e non cinetici" sono quelli in cui l'IA è impiegata per operazioni cibernetiche sia di difesa sia di attacco con obiettivi non cinetici. Quelli "conflittuali e cinetici" si riferiscono all'integrazione dei sistemi IA in operazioni di combattimento: si va dall'impiego di sistemi IA per l'identificazione delle minacce al loro uso in sistemi d'arma autonomi letali.

Nel resto del libro, il capitolo 1 fornisce le basi dell'analisi, delineandone la metodologia e l'ambito. Il capitolo 2 descrive cinque principi fondamentali dell'etica dell'IA nella difesa e una metodologia per implementarli. I sei capitoli successivi si concentrano su differenti usi dell'IA nella difesa, spaziando dall'analisi di intelligence coadiuvata dall'IA alla guerra cibernetica, alla deterrenza cibernetica e ai sistemi d'arma autonomi. Chi fosse interessato a usi specifici dell'IA nella difesa può passare dall'uno all'altro fra questi ultimi capitoli, ma sarebbe bene che leggesse comunque i primi due.

Prima di concludere questa prefazione, devo affrontare una domanda che probabilmente è sorta in chi legge il sottotitolo di questo libro: perché preoccuparsi dell'etica dell'IA nella difesa? La domanda mi è stata posta molte volte, mentre discutevo di questo lavoro con studiosi, politici ed esperti della difesa. Spero che l'analisi offra motivi sufficienti per giustificare una riflessione profonda e una discussione articolata sull'etica dell'IA nella difesa. Tuttavia, visto che non affronto direttamente la risposta a questa domanda nel resto del libro, colgo l'occasione per parlarne qui, sia pur brevemente.

La domanda viene posta in genere da persone che appartengono all'uno o all'altro di due campi opposti. Il primo è quello di chi persegue l'uso dell'IA nella difesa e percepisce l'etica come un ostacolo all'adozione di questa tecnologia, perché potrebbe ridurre il vantaggio sugli avversari che possono acquisire e utilizzare l'IA senza preoccuparsi delle implicazioni etiche. Chi appartiene a questo campo abbraccia spesso una concezione realistica della politica internazionale. Il pacifismo, invece, motiva chi appartiene al campo opposto. Qui l'assunto è che non possa derivare alcun bene dalla difesa e, a fortiori, dall'uso dell'IA nella difesa, perciò non è necessaria alcuna riflessione etica, perché qualsiasi uso di questa tecnologia nella difesa dovrebbe essere proibito. Secondo questa posizione, questo tipo di analisi etica non è necessario ed è problematico, poiché può finire per legittimare l'uso dell'IA nella difesa.

Sono convinta che gli assunti alla base di entrambi i campi non siano difendibili. Da un lato, le società democratiche liberali si impegnano in una guerra *giusta* quando la fanno per difendersi. È impensabile entrare in un conflitto per difendere quelle società senza alcun riguardo per l'etica, e quindi per i valori e i diritti fondamentali che si vogliono tutelare. Per le democrazie liberali, questo significherebbe sconfiggere sé stesse. Il realismo è una posizione pericolosa, perché lascia che sia l'avversario a definire le condizioni secondo cui una società democratica liberale rispetta e mette in pratica i suoi valori e diritti fondamentali. D'altro lato, le posizioni pacifiste restano, secondo me, troppo idealistiche. Concordo: la guerra è un male assoluto da evitare, ma è un male in cui gli esseri umani sono stati regolarmente coinvolti nel corso della storia – e purtroppo continuano a esservi coinvolti. Se l'aggressione di uno Stato a un altro Stato o popolo rimane ingiustificabile, anche il diritto di uno Stato a difendersi e opporre resistenza a un aggressore rimane innegabile. È

fondamentale una considerazione seria dell'etica della guerra per evitare che quel diritto porti alla perfidia, e quindi all'atrocità, in guerra (Lippert-Rasmussen, 2014).

Chi legge e non fosse d'accordo con queste motivazioni (politiche) a sostegno dell'urgenza di un'etica dell'IA nella difesa potrebbe apprezzare la concezione che propone Walzer quando collega l'etica della guerra alla strategia. Secondo Walzer, entrambe offrono un linguaggio della giustificazione; vale a dire, sia il ragionamento etico sia quello strategico si preoccupano dell'identificazione di rischi e opportunità e di definire una condotta per mitigare i primi e sfruttare le seconde. Come scrive Walzer:

> Il teorico morale […] deve constatare che le sue regole vengono spesso violate o ignorate – e, cosa ancor più importante, che a chi partecipa alla guerra le sue regole non appaiono rilevanti in rapporto alla eccezionalità delle condizioni in cui si trova ad agire. Nonostante ciò, però, egli non deve rinunciare alla sua interpretazione della guerra in quanto azione umana, intenzionale e premeditata, dei cui effetti qualcuno va pur ritenuto responsabile. Nel momento in cui si confronta con i crimini commessi nel corso della guerra, o con il crimine costituito dall'aver scatenato una guerra d'aggressione, egli va alla ricerca dei suoi autori. E non è solo in questa ricerca: uno degli aspetti più eclatanti della guerra, che la distingue da ogni altro flagello del genere umano, è che uomini e donne in essa coinvolti oltre ad esserne le vittime vi partecipano anche attivamente. Tutti noi siamo portati a ritenerli responsabili di ciò che fanno [… Potremmo] dire di un generale che non ha avuto alcuna difficoltà a prendere una decisione (realmente) difficile che non ha colto la valenza strategica della propria posizione o che è stato imprudente e insensibile al pericolo. E potremmo spingerci oltre, nel caso del generale, arrivando a sostenere che un uomo simile non avrebbe alcun diritto di combattere o di guidare altri soldati in battaglia, che dovrebbe […] preoccuparsi del pericolo che ciò comporta e adottare misure per evitarlo. *Ancora una volta la situazione è identica nel campo delle decisioni morali: soldati e statisti dovrebbero conoscere i pericoli della crudeltà e dell'ingiustizia, preoccuparsi di essi e fare in modo di evitarli.* (1977, pp. 19-21, corsivo mio)

Questo non è meno vero quando si considerano l'uso dell'IA nella difesa e il rischio che gli esseri umani possano basarsi su queste tecnologie per perpetrare crudeltà e ingiustizia nella condotta bellica. La storia delle guerre del Novecento ci ricorda l'incredibile livello di atrocità che siamo capaci di infliggerci gli uni gli altri e come abbiamo saputo far leva sulla tecnologia a questo fine. Considerando che l'IA è progettata per massimizzare (e persino superare) le capacità umane, l'urgenza e la necessità di un'analisi etica che ne guidi l'uso nella difesa dovrebbero apparire evidenti a tutti.

1. Alcuni dei capitoli di questo libro sono apparsi in origine come articoli o saggi per convegni. L'elenco completo di queste pubblicazioni si può trovare al termine dei ringraziamenti.

2. *Wales Summit Declaration*, NATO, pubblicato il 5 settembre 2014, https://www.nato.int/cps/en/natohq/official_texts_112964.htm (ultimo accesso agosto 2024).

RINGRAZIAMENTI

Sarebbe stato impossibile portare a termine questo libro senza il sostegno di colleghi, amici e familiari. Ho la fortuna di avere un enorme debito di gratitudine con tutti loro. Inizierò con i colleghi con cui ho lavorato a più stretto contatto nel corso degli anni. Sono profondamente grata perché hanno condiviso con me le loro idee, le intuizioni, le domande e l'entusiasmo per cercare di capire le trasformazioni legate alla rivoluzione digitale e l'impegno per contribuire a guidare queste trasformazioni, sia pure in minima parte, con il nostro lavoro. Rachel Azafrani, Alexander Blanchard, Josh Cowls, Prathm Juneja, Jakob Mökander, Jess Morley, Carl Öhman, Huw Roberts, David Sutcliffe, Christopher Thomas, Andreas Tsamados, Vincent Wang, David Watson e Marta Ziosi hanno contribuito a migliorare in modi significativi le mie idee sull'etica dell'IA nella difesa.

Ho discusso i contenuti di questo libro con molti colleghi nel mondo accademico, in quello del governo e in quello della difesa. Sono grata per tutti i suggerimenti, i case study, le critiche e i riferimenti che i colleghi hanno condiviso con me nel corso degli anni; alcuni sono stati cruciali per migliorare parti fondamentali del mio lavoro. Sono particolarmente grata a Al Banks, David McNeish e Leila Kleineidam per il loro impegno a discutere di etica dell'IA nel campo della difesa, e per il tempo che hanno dedicato a leggere e commentare alcuni testi confluiti in questo libro. Devo anche riconoscere il sostegno, i suggerimenti e le parole di incoraggiamento che mi hanno dato in tutti questi anni Rebecca Eynon, George Lucas, Ties Nijssen e Selmer Bringsjord. Mi hanno aiutata a superare i momenti di incertezza che inevitabilmente si presentano in un progetto di questo genere. Sono anche grata a Peter Ohlin per averci creduto abbastanza da decidere di pubblicare la versione inglese; a Raffaello Cortina e Albertine Cerutti per la pazienza con cui hanno seguito la pubblicazione della versione italiana.

Parti di questo libro sono state pensate inizialmente come articoli di ricerca o capitoli di libri (come è indicato nell'elenco dei riferimenti bibliografici), e sono il risultato di diversi progetti di ricerca che ho guidato negli scorsi decenni. Sono grata ai finanziatori per il loro sostegno, in particolare alla Marie Skłodowska-Curie Action, al John Fell Fund dell'Università di Oxford, al NATO Centre of Excellence on Cooperative Cyber Defence, allo Alan Turing Institute di Londra, al Defence Science and Technology Laboratory del ministero della Difesa del Regno Unito. Alcuni di questi finanziamenti sono stati concessi con ammirevole lungimiranza, molto prima che l'etica dell'IA, in particolare nella difesa, fosse un ambito di ricerca consolidato. Sorvolo su tutte le proposte che non sono state finanziate e gli articoli che sono stati rifiutati: sia pure con poco entusiasmo, sono grata anche per questi insuccessi. Ogni volta, mi hanno consentito di migliorare il mio progetto e di approfondire le mie idee.

Ringrazio per il sostegno incrollabile e la premura un gruppo notevole di persone amiche: Antonella Giglio, Dominik Aschbrenner, Elena Brenna, Elisa Sacchi, Francesco Fermani e Giorgia Brambilla Pisoni. Poter contare sul loro affetto, trovare in loro sparring partner incoraggianti e onesti nel districare i miei pensieri, rompere la bolla accademica e creare bei ricordi, è stato fondamentale per conservare un po' di equilibrio durante la scrittura di questo libro. Sono grata anche alla mia famiglia: da loro ho imparato la determinazione e l'ostinazione, che sono state fondamentali per portare a termine questo libro. Trovo in loro una fonte di sostegno pragmatico, a volte severo, che mi mantiene con i piedi per terra. La mia gratitudine, profonda e duratura, si estende a Luciano Floridi e Kia Nobre: trovare le parole giuste per esprimere il mio affetto e il mio apprezzamento per loro è difficile, perché sono figure da cui ho tratto grande ispirazione. Il loro lavoro, la genialità delle loro intuizioni, la generosità, l'integrità, l'ironia, l'energia illimitata, la brillantezza, l'amicizia, l'affetto, il sostegno e, soprattutto, il loro entusiasmo genuino per la ricerca sono una fonte costante di motivazione per me. Rimane un po' di imbarazzo, perché, nonostante la forza che ha avuto e continua ad avere il loro esempio, riesco a seguirlo solo in parte.

Nonostante tutto l'aiuto ricevuto, sono consapevole che nelle pagine che seguono potrebbero esserci ancora degli errori: la responsabilità in quel caso è solo mia.

I capitoli di questo libro sono adattati dalle pubblicazioni seguenti:

- capitolo 1: Taddeo et al., 2022; Taddeo et al., 2021;
- capitolo 2: Taddeo et al., 2021; Taddeo, Blanchard, Thomas, 2023;
- capitolo 3: Blanchard, Taddeo, 2023;
- capitolo 4: Taddeo, 2014a, 2016b, 2017a, 2018a/b/c;
- capitolo 5: Taddeo, 2017b, 2018a, 2018c;
- capitolo 6: Blanchard, Taddeo, 2022b;
- capitolo 7: Taddeo, Blanchard, 2022;
- capitolo 8: Blanchard, Taddeo, 2022a, 2022b, 2022c.

ABBREVIAZIONI PIÙ FREQUENTI

AIA:	IA per l'analisi di intelligence (*augmented intelligence analysis*).
AWS:	sistemi d'arma autonomi (*autonomous weapon systems*).
DL:	*deep learning* (apprendimento profondo).
DoD:	Dipartimento della difesa (US Department of Defense).
EB:	comitato etico (*ethics board*).
HMT:	team umani-macchine (*human-machine teaming*).
HMT-AI:	team umani-macchine che comprendono sistemi IA (*human-machine teaming including AI systems*).
IA:	intelligenza artificiale.
ICRC:	International Committee of the Red Cross.
IHL:	diritto umanitario internazionale (*international humanitarian law*).
LAWS:	sistemi d'arma autonomi letali (*lethal autonomous weapon systems*).
LdA:	livello di astrazione.
LLM:	modelli linguistici di grandi dimensioni.
ML:	*machine learning* (apprendimento automatico).
MoD:	ministero della Difesa (UK Ministry of Defence).
RAI:	intelligenza artificiale responsabile (*responsible artificial intelligence*).
RRI:	ricerca e innovazione responsabile (*responsible research and innovation*).

1

LE BASI PER UN'ETICA DELL'IA NELLA DIFESA

1.1 INTRODUZIONE

Gli impieghi dell'IA nella difesa sono di vario tipo: vanno dall'analisi predittiva con algoritmi di *machine learning* (ML) per migliorare la gestione di rifornimenti e apparecchiature all'analisi di grandi quantità di dati a supporto dell'intelligence e dei processi decisionali concernenti l'applicazione della forza, all'uso della forza stessa. Dai tweet ai carri armati, l'IA ora è una risorsa fondamentale nella difesa e perciò è oggetto di una corsa globale per il suo sviluppo e uso (Taddeo, Floridi, 2018a). Per esempio, dal 2014, il Centro di controllo nazionale della difesa russo utilizza l'IA per individuare minacce online. Nel 2017 il governo cinese ha definito il suo Next Generation AI Development Plan, di cui l'implementazione militare dell'IA sul campo di battaglia e nel cyberspazio è una parte cruciale. Regno Unito e Stati Uniti, Francia e Australia hanno pubblicato strategie di difesa nazionale centrate sull'uso dell'IA. La NATO ha lavorato e lavora per rafforzare l'innovazione nella difesa, e l'IA ricopre un posto fondamentale nella sua strategia. L'IA è già utilizzata anche nei conflitti. Nel 2023, Israele ha impiegato sistemi IA per l'identificazione di obiettivi umani a Gaza, secondo quanto è stato riferito,[1] ed entrambi i fronti della guerra in Ucraina si servono ampiamente dell'IA.[2]

L'uso dell'IA per la difesa nazionale pone problemi etici importanti, in cui si combinano i rischi etici relativi al suo stesso impiego – per esempio, la possibilità di violazioni dei diritti umani, la riduzione del controllo, la mancata attribuzione della responsabilità morale, la svalutazione delle competenze di agenti umani e l'erosione della loro autodeterminazione (Yang et al., 2018) – e quelli che derivano dall'uso della forza nel combattimento, come la violazione della dignità umana e dei principi della Teoria della Guerra Giusta.

Dato il ventaglio delle possibili applicazioni e l'insieme, ampio e complesso, dei rischi etici che si devono affrontare, risulta difficile sviluppare un'analisi etica coerente e sistematica dell'IA nella difesa. I problemi iniziano già con la definizione di IA (perché questo termine si riferisce a un insieme di approcci, metodi e modelli) e proseguono quando si cerca di determinare il raggio d'azione dell'analisi etica. Per affrontare questi problemi, si potrebbe pensare di sviluppare una tassonomia delle questioni etiche relative all'IA nella difesa, ma sarebbe un'impresa irrealizzabile e comunque di scarso valore: gli sviluppi nelle tecnologie dell'IA e le loro applicazioni in nuovi ambiti renderebbero presto la tassonomia obsoleta. Allo stesso tempo, l'identificazione dei problemi etici varia con il punto di vista preso in considerazione. Per esempio, alcuni problemi etici dell'IA sono intrinseci al processo di progettazione e sviluppo, mentre altri emergono all'interno di domini specifici e con specifiche finalità di utilizzo. Questo rende difficile identificare il livello giusto di analisi. Se si trascurano le caratteristiche dell'ambito della difesa e le finalità d'uso, le analisi rischiano di essere troppo generiche per offrire una guida concreta. D'altra parte, le analisi etiche che si limitano all'uso dell'IA in un campo specifico devono comunque rimanere coerenti con l'insieme più generale dei valori che stanno alla base delle nostre società. Per evitare questi limiti, o come minimo per esserne ben consapevoli, bisogna chiarire subito gli assunti relativi all'ambito dell'analisi e alla sua metodologia. Questo è lo scopo del capitolo.

Darò una definizione di IA e di alcune delle caratteristiche tecniche fondamentali di questa tecnologia che sono rilevanti per un'analisi etica del suo uso nella difesa, concentrandomi in particolare, nel paragrafo 1.2, sul "problema della predicibilità", che apre la porta ai rischi etici più noti dell'IA. È crescente la preoccupazione che, per prendere decisioni la cui posta è molto alta, l'uso di sistemi IA poco prevedibili possa avere esiti gravi (Holland Michel, 2020b), che vanno dai rischi di sicurezza per infrastrutture critiche ai rischi per i diritti e il benessere degli individui, all'escalation dei conflitti o a ricadute diplomatiche e violazioni dei principi della Teoria della Guerra Giusta. Il "problema della predicibilità" è quindi centrale in questo libro, ed è il motivo per cui lo introduco qui. Poi, nel paragrafo 1.3, presenterò la metodologia dei *livelli di astrazione* (LdA) (Floridi, 2008). Questa metodologia ci consentirà di determinare l'ambito dell'analisi che delineerò nel paragrafo 1.4, dove descriverò le tre

categorie di usi dell'IA nella difesa, che definiscono la struttura del libro. Le conclusioni sono contenute nel paragrafo 1.5.

1.2 L’IA E IL PROBLEMA DELLA PREDICIBILITÀ

L’IA si può definire in modi diversi, per esempio concentrandosi sull’automazione del comportamento intelligente o sulla progettazione di agenti intelligenti e di modelli computazionali del comportamento umano. Ai fini di questo libro, trascureremo gli aspetti tecnici specifici di un sistema (per esempio, se si tratti di un sistema statistico o sub-simbolico) e ci focalizzeremo invece solo sulle caratteristiche specifiche dei sistemi IA che danno luogo a sfide etiche. Nel resto del libro, quando parlerò di IA, mi riferirò a “una risorsa crescente di *agency* interattiva, *autonoma* e *ad apprendimento*, che può essere utilizzata per svolgere compiti che, per essere portati effettivamente a termine, richiederebbero altrimenti l’intelligenza umana” (Floridi, Cowls, 2019, corsivo mio). La combinazione di agency, autonomia e capacità di apprendimento è l’aspetto cruciale, perché è alla base sia degli usi benefici dell’IA sia di quelli problematici. È anche l’origine del problema della predicibilità, ossia la limitata sicurezza con cui si può rispondere alla domanda: “Che cosa farà un sistema IA?”.

L’impredicibilità dei sistemi non è una novità. Si danno sistemi impredicibili in matematica e fisica, e l’esistenza di limiti alla capacità di prevedere i comportamenti di sistemi artificiali è stata dimostrata formalmente già negli anni Cinquanta (Rice, 1956; Moore, 1990; Musiolik, Cheok, 2021). Wiener e Samuel hanno discusso la predicibilità dei sistemi IA in un famoso scambio nel 1960 (Wiener, 1960; Samuel, 1960), in cui il primo attribuiva la non predicibilità alle capacità di apprendimento di quei sistemi e notava che, “poiché le macchine apprendono, possono sviluppare strategie impreviste a velocità che lasciano spiazzati i loro programmatori” (Wiener, 1960, p. 1355). Gli sviluppi delle ricerche sull’IA hanno dimostrato che Wiener aveva ragione. Pensiamo, per esempio, al *reward hacking*, che è stato indicato nella letteratura come una delle cause del problema della predicibilità. In questo caso,

> agenti autonomi ottimizzano la funzione di ricompensa [data dai progettisti] […]. Quando progettiamo la ricompensa, potremmo pensare a qualche specifico scenario di addestramento, e assicurarci che la ricompensa porti al comportamento giusto in quegli

> scenari. Inevitabilmente, gli agenti incontrano *nuovi* scenari (per esempio, nuovi tipi di terreno) nei quali l'ottimizzazione della medesima ricompensa può condurre a un comportamento indesiderato. (Hadfield-Menell et al., 2020, p. 1)

La predicibilità dei sistemi IA oggi tende a essere vista o come un problema tecnico – una conseguenza diretta delle caratteristiche dei sistemi IA (International Committee of the Red Cross, 2019; Boulanin et al., 2020; Defense Innovation Board, 2019) – o come un problema operativo derivante dall'interazione del sistema con l'ambiente in cui è utilizzato (International Committee of the Red Cross, 2019; Docherty, 2020).

La predicibilità tecnica di un sistema IA viene valutata nei termini del grado di coerenza fra i suoi comportamenti passati, correnti e futuri (Holland Michel, 2020a). Gli aspetti fondamentali monitorati in questo caso sono *cambiamenti di dati* e *concept shift* (slittamento concettuale di un modello IA); quanto spesso e per quanto tempo gli output di un sistema sono corretti; e se il sistema può essere esteso in modo da gestire correttamente dati divergenti da quelli usati in fase di sviluppo (Boulanin et al., 2020; Collopy, Sitterle, Petrillo, 2020; Defense Innovation Board, 2019). La predicibilità tecnica dipende anche da caratteristiche come l'interpretabilità, la trasparenza, l'esplicabilità e la robustezza dei sistemi IA (Holland Michel, 2020a; Rudin, Wang, Coker, 2020). Vale la pena di sottolineare che, nei sistemi IA, una predicibilità limitata non equivale a una robustezza limitata: la robustezza si riferisce alla capacità di un sistema IA di produrre risultati corretti anche quando riceve in ingresso dati non corretti, mentre la predicibilità si riferisce alla probabilità che il sistema si comporti come atteso, indipendentemente dalla correttezza dei suoi esiti. Tuttavia, quando si parla di predicibilità tecnica, solitamente si concentra l'attenzione sull'errore o sulla manipolazione del sistema, ragion per cui la distinzione fra esiti imprevisti ma corretti ed esiti non previsti e sbagliati spesso si perde. In questo caso, nessuno dei sistemi è prevedibile, ma quello dagli esiti corretti è perlomeno robusto.

La predicibilità operativa si riferisce alla misura in cui le azioni di un sistema possono essere previste, quando questo è utilizzato in un ambiente specifico. In tal senso, "tutti i sistemi autonomi presentano un certo grado di impredicibilità operativa intrinseca, anche se non falliscono o se gli esiti delle loro singole azioni possono essere ragionevolmente previsti" (Holland Michel, 2020b, p. 5). Sulla predicibilità operativa incide un

ampio insieme di variabili: le funzionalità tecniche del sistema, le caratteristiche del contesto di utilizzo, le interazioni con altri sistemi, il grado di comprensione, da parte dell'operatore, del funzionamento del sistema e, nel campo della difesa, il comportamento degli avversari. Questi fattori possono variare e interagire in modi complessi, il che rende difficile prevedere tutti i possibili output di un sistema IA e i loro effetti.

Il problema della predicibilità è multidimensionale, e per questo la distinzione fra predicibilità tecnica e operativa di un sistema IA non è effettivamente plausibile nella pratica, perché tutti i fattori tecnici e operativi contribuiscono a determinare il comportamento del sistema. Altrove (Taddeo et al., 2022) ho sostenuto che il problema della predicibilità si può inquadrare nel migliore dei modi come un intervallo di output che variano per livello di predicibilità, essendo il risultato della combinazione e dell'interazione di aspetti tecnici, di sicurezza e operativi. A questo fine, ho definito i limiti inferiore e superiore del problema della predicibilità:

> Come *minimo*, dato uno scenario ideale in cui si può dare per scontato o per rilevato che non vi sono stati errori nelle fasi di progettazione e di sviluppo, una volta messo in esercizio un sistema IA può comunque produrre esiti corretti (e tuttavia non desiderati), non prevedibili al momento in cui è stato iniziato il suo utilizzo.
>
> Al *massimo*, dati i molti aspetti dei processi di progettazione, sviluppo e messa in esercizio dei sistemi IA, l'opacità di questi sistemi, la loro capacità di adattamento e le possibili complessità dell'ambiente di utilizzo, non è possibile rendere conto né di tutte le fonti di errore e di manipolazione di un sistema, né di tutti i possibili comportamenti emergenti (positivi o no) di un sistema IA che possono essere determinati da quegli errori. (*Ibidem*, p. 15)

È importante chiarire una cosa, prima di prendere in considerazione più dettagliatamente alcune delle cause fondamentali del problema della predicibilità: l'impredicibilità di un sistema IA non è sconfinata (*boundless*), ma è limitata dalle *affordances* del sistema, cioè dall'insieme delle specifiche hardware e software che determinano la gamma delle possibili azioni di una macchina. Per esempio, un sistema IA progettato e sviluppato per distinguere immagini di cavalli da quelle di cani può essere impredicibile per quanto riguarda il modo in cui gestisce gli input visuali e la selezione finale delle immagini, ma non c'è timore che il sistema sviluppi un comportamento imprevisto al di fuori delle sue *affordances* e produca un nuovo tipo di risultato, per esempio che disegni l'immagine di un cavallo o di un cane o che nitrisca. Ne segue che, al crescere di livello e

complessità delle *affordances* di un sistema, cresce anche la gamma dei comportamenti impredicibili che il sistema potrà sviluppare una volta messo in uso.

1.2.1 Team umani-macchine

All'estremità inferiore dello spettro prospettato dalla definizione di minima e massima predicibilità data sopra, possiamo immaginare un sistema realizzato grazie a un processo impeccabile di progettazione e sviluppo, che tuttavia mostra comunque alcuni esiti imprevisti, in conseguenza delle sue capacità di apprendimento e delle interazioni con l'ambiente o della modalità d'uso. Mi concentrerò qui sulla modalità d'uso perché, contrariamente all'ambiente, questa può essere progettata in modo da mitigare i rischi di esiti imprevisti. Al contempo, la modalità d'uso dei sistemi IA diventa un aspetto fondamentale da prendere in considerazione quando si passa dall'uso dell'IA come strumento alla sua integrazione come agente artificiale in un team di lavoro.

Nei team umani-macchine che comprendono sistemi IA (*human-machine teaming including AI systems*, HMT-AI), fattori non solo tecnici ma anche culturali, etici, giuridici e cognitivi possono contribuire tutti al verificarsi di esiti imprevisti (Andras et al., 2018; Chopra, Singh, 2018; Ehsan, Riedl, 2020; Makarius et al., 2020; NIST, 2022). Gli HMT-AI segnano una svolta rispetto a modi precedenti di utilizzo dell'IA, che prevedevano una chiara divisione del lavoro fra esseri umani e agenti artificiali, si basavano su bassi livelli di automazione e assegnavano l'elaborazione di più fonti di informazione solo agli esseri umani (Shaw et al., 2010; Walliser et al., 2019; Woods, Patterson, Roth, 2002). Gli HMT-AI si concentrano sulla produzione di sistemi di intelligenza congiunta, in cui i compiti sono distribuiti fra esperti umani e sistemi IA in modo da creare processi agili e facilitare capacità emergenti (O'Neill et al., 2020). Gli HMT-AI attualmente sono impiegati in vari campi (Lavin et al., 2021; Scherrer et al., 2022), fra cui la logistica (Stowers et al., 2021), le squadre di ricerca e salvataggio in ambiente urbano, le équipe chirurgiche avanzate (You, Robert, 2016) e la cybersicurezza (Stevens, 2020), dove i ricercatori combinano l'esperienza e l'intuizione degli esperti con tecniche di ML per creare un sistema in grado di rilevare attacchi imprevisti e di difendersi da quegli attacchi (Veeramachaneni et al., 2016). Anche le operazioni di

difesa si basano su HMT-AI (Konaev, Chahal, 2021; Lopez, 2022) nel supporto ai processi decisionali, che comprendono l'analisi di dati grezzi e l'analisi predittiva in operazioni di intelligence per consentire l'esame di dati multidimensionali che altrimenti rimarrebbero inutilizzati (National Academies of Sciences, Engineering, and Medicine, 2022).

La fiducia degli esseri umani negli agenti artificiali è una componente chiave e problematica degli HMT-AI, che può anche aggravare il problema della predicibilità, se gli agenti umani hanno livelli inappropriati di fiducia nell'agente artificiale. Gli operatori possono fidarsi poco dell'agente artificiale e perciò supervisionare in maniera eccessiva le sue azioni. Questo crea costi opportunità, poiché gli agenti umani sprecano tempo e risorse nella supervisione dell'agente artificiale, anziché svolgere i compiti che l'IA non può affrontare. La bassa fiducia nell'IA ha poche conseguenze per il problema della predicibilità. Non succede lo stesso con l'eccesso di fiducia, quando gli agenti umani delegano troppo all'agente artificiale ed esercitano troppo poca supervisione, con il risultato di correre rischi non necessari. In contesti di decisioni ad alto rischio, l'eccesso di fiducia aggrava i rischi dell'impredicibilità e può portare a esiti avversi, fra cui rischi elevati per gli esseri umani che interagiscono con l'agente artificiale. Per esempio, uno studio del 2016 ha descritto un esperimento, condotto con 42 volontari in un'emergenza di incendio simulato, con una guida robotica il cui compito era portarli al sicuro (Robinette et al., 2016). Quasi tutti i partecipanti hanno seguito ciecamente il robot, che ha commesso vari errori fatali, di cui non ha dato né una spiegazione né un avvertimento.

L'eccesso di fiducia può generare anche una dinamica *trust and forget*, ovvero "fidati e dimenticatene" (Taddeo, 2017c), in cui l'essere umano ha il massimo livello di fiducia nell'agente artificiale, non esercita una supervisione della sua performance e non rileva le sue azioni (potenzialmente erronee), non tenendo conto delle sue capacità e dei suoi limiti, ma accettandone acriticamente gli esiti. Questo eccesso di fiducia può essere il risultato di un bias dell'automazione nell'IA, cioè della tendenza degli agenti umani a fidarsi eccessivamente degli output dell'IA (Goddard, Roudsari, Wyatt, 2012). Come sostiene Struß, questo bias e il rischio dell'eccesso di fiducia diventano più problematici con il crescere della complessità dei sistemi IA, perché gli agenti umani e quelli artificiali negli HMT-AI hanno processi decisionali diversi ed è possibile che l'agente

umano non sia in grado di vagliare o comprendere come l'agente artificiale raggiunge le sue decisioni, ma si basi comunque su quelle credendo che la macchina esegua sempre i suoi compiti in modo corretto.[3]

Programmi di addestramento, però, possono aiutare gli agenti umani a calibrare le loro aspettative e sviluppare una comprensione più accurata del comportamento generale del sistema, evitando problemi di interfaccia e di fiducia. Tuttavia, questi programmi di addestramento richiedono concetti, metodi e standard nuovi per essere adattati con successo agli HMT-AI (Laird, Ranganath, Gershman, 2019; Lavin et al., 2021; National Academies of Sciences, Engineering, and Medicine, 2022). Purtroppo, la letteratura sugli HMT-AI si basa ancora su strutture sviluppate per i più tradizionali HMT, dotati di livelli inferiori di automazione e di minori capacità di apprendimento. Questo costituisce un problema evidente, dato che le caratteristiche e le dinamiche degli HMT non corrispondono perfettamente a quelle degli HMT-AI, in particolare per il modo in cui i ruoli e gli obiettivi vengono assegnati dinamicamente in nuovi contesti d'impiego, per il modo in cui sono sviluppate le rappresentazioni condivise e per il modo in cui viene attribuita la responsabilità.

I metodi specifici di addestramento richiesti dagli HMT-AI restano inesplorati (McNeese et al., 2021; O'Neill et al., 2020). Nel progettare nuovi modi di addestramento è fondamentale includere regolarmente incertezze e perturbazioni, per aiutare *sia* gli agenti umani *sia* quelli artificiali a costruire rappresentazioni complete dei criteri decisionali di ciascuno (Niu, Paleja, Gombolay, 2021; Shih et al., 2021). Per esempio, studi sulla fiducia nei robot guida in contesti di emergenza (Robinette, Howard, Wagner, 2017) hanno evidenziato come l'esperienza preliminare di comportamenti errati da parte dei robot possa generare benefici significativi per gli esseri umani che sono parte di un HMT-AI, migliorando la consapevolezza e la gestione del rischio in situazioni reali. Allo stesso tempo, i modelli mentali perfezionati dagli agenti umani durante le sessioni di addestramento possono essere utilizzati per migliorare le performance di agenti artificiali o l'interfaccia che facilita la comunicazione fra gli agenti, creando un *feedback loop* (ciclo di retroazione) (Klamm et al., 2019). Il valore di tutto questo è stato esaminato per gli HMT-AI in giochi di strategia in tempo reale (A. Anderson et al., 2020), in *war games* militari (Schwartz et al., 2020), in team di volo

autonomo (Tossell et al., 2020) e nella cybersicurezza (Buchanan, Imbrie, 2022; Ding et al., 2019; Gomez, Mancuso, Staheli, 2019).

1.2.2 *Machine learning*

Se passiamo ora al limite superiore dello spettro del problema della predicibilità dobbiamo considerare gli esiti imprevisti risultanti da combinazioni di errori nelle fasi di progettazione e sviluppo. Qui sono determinanti tre fonti di errore: i modelli per il ML, la cura dei dati, i processi di progettazione e sviluppo.

Uno dei problemi principali associati ai modelli ML dalle prestazioni migliori (per esempio, reti neurali e alberi di decisione potenziati) è che la loro complessità rende difficile valutare se i modelli effettuano generalizzazioni appropriate per dati che sono al di fuori delle distribuzioni del dataset di addestramento. La *confidence* del modello (il grado di certezza con cui il modello effettua una previsione o una classificazione) è l'approccio più comune nel moderno ML per trattare l'incertezza associata alla generalizzazione, valutando le differenti incertezze che caratterizzano il modello e il suo ambiente operativo (Hüllermeier, Waegeman, 2021). La *confidence* del modello, però, spesso non è robusta da un punto di vista statistico. Le *deep neural networks* (o reti neurali profonde), per esempio, si sono dimostrate troppo sicure di sé, e per questo possono condurre a errori per un eccesso di fiducia o nascondere accidentalmente attacchi contro il modello (ENISA, 2020). I livelli di *confidence*, perciò, devono essere adeguati agli output, il che complica (e può perturbare) i processi successivi.

Anche per i modelli IA dalle prestazioni migliori, gli esiti dell'addestramento non sono necessariamente indicativi delle capacità di un sistema nel mondo reale, dove le condizioni d'uso possono divergere da quelle di sviluppo, e i dati ricadere al di fuori delle distribuzioni del dataset di addestramento. In letteratura sono riportati spesso casi di modelli di *deep neural networks* che non riescono a generalizzare in modo appropriato al di fuori delle condizioni di addestramento (Nguyen, Yosinski, Clune, 2015; Athalye et al., 2018). Nella visione automatica, per esempio, è difficile analizzare immagini in presenza di un contesto rumoroso o quando pixel o luci estranee provocano una confusione contestuale. I sistemi IA sono suscettibili a cambiamenti anche di piccola

entità, addirittura a livello di pixel, e piccole variazioni possono far sì che un sistema identifichi erroneamente delle strisce pedonali come autobus scolastici (Nguyen, Yosinski, Clune, 2015). È stato dimostrato che questi limiti sono sfruttabili in operazioni multidominio abilitate dall'IA in campo militare (Jia et al., 2022; Savas et al., 2020).

1.2.3 Cura dei dati

La cura dei dati è un passo fondamentale nello sviluppo di sistemi IA e il *data labelling* (o etichettatura dei dati) è un aspetto chiave per quanto riguarda il problema della predicibilità. Le *data labels* attribuiscono un significato ai dati di addestramento e consentono alla macchina di apprendere. Esistono metodologie diverse per il *data labelling*, ma tutte hanno limiti importanti che possono portare a esiti imprevedibili del sistema. Per esempio, chi etichetta i dati può riprodurre bias (Bekele, Narber, Lawson, 2017; Bekele et al., 2018), creando dataset di addestramento sbilanciati che incideranno sulla performance del sistema IA e potranno portare a esiti imprevisti. Altre forme di *data labelling*, come quella basata sul consenso, possono migliorare la qualità complessiva ma a costi superiori. In qualche caso, ci si può affidare a un *data labelling* sintetico, che però richiede capacità di elaborazione enormi e ha un elevato potenziale di errore (IBM, 2021).

Il *data cleaning* (la pulizia dei dataset) è un'altra forma di cura: si eliminano dal dataset i doppioni, le caratteristiche non informative e i "fuoriclasse" (Tobin, 2022), al fine di rimuovere il rumore e migliorare le performance del modello. C'è il rischio, comunque, che in questo modo scompaiano anche punti di dati significativi, il che può portare a esiti non voluti nella fase applicativa se il dataset risultante è stato privato di informazioni importanti, utili per testare il comportamento del modello. Quando i dataset sono costruiti partendo da più fonti di input, alcune delle quali possono essere fra loro incompatibili, emerge il problema del *data commingling* (o commistione dei dati). Questo accade, per esempio, quando si utilizzano sensori differenti che non sono stati calibrati o normalizzati in modo da produrre gli stessi valori: così si ottengono dataset incoerenti, incompleti e non accurati, nonché risultati inaffidabili.

Il *data shift* si riferisce a un cambiamento nella distribuzione dei dati in conseguenza di fattori esterni (Sarantitis, 2020) che può portare a risultati

erronei. Per costruire un modello IA è necessario identificare le relazioni prevedibili fra variabili di input e variabili target, sulla base dell'assunto che a parità di distribuzione dei dati si ottengono risultati simili. Nel mondo reale, però, fattori imprevedibili possono modificare gli input, la qualità del dataset, la raccolta dei dati (per esempio, la frequenza delle operazioni di *polling*, la verifica ciclica di input e output) o addirittura gli schemi sottostanti che danno forma alle relazioni fra dati di input e di output (*ibidem*).

Oltre alla possibilità di introdurre errori non voluti negli output di un sistema IA, la cura dei dati pone due sfide operative importanti per quanto riguarda la predicibilità del comportamento di un sistema. La prima è il *pay-off* operativo (lo sforzo fatto rispetto all'efficienza risultante) della cura dei dati; la seconda emerge in caso di assenza di standard e meccanismi automatici per migliorare la qualità dei dati. Alcune dimensioni fondamentali della qualità dei dati sono la completezza, l'accuratezza, l'unicità, la tempestività, la coerenza e la validità. Queste dimensioni e la loro importanza relativa, però, possono variare in funzione dei diversi contesti d'uso e delle finalità correlate. Ancora non esiste una definizione degli standard di qualità e dei meccanismi di governance per dati non strutturati, il che è problematico, perché sono questi i tipi di dati che si utilizzano in generale nei modelli IA. Come ha evidenziato l'EU Agency for Fundamental Rights (FRA, 2019), questa mancanza di specifiche di qualità e di una guida porta all'assenza di standard, strumenti e meccanismi efficienti[4] per valutare in modo robusto i dati e verificare se sono adeguati allo scopo che ci si prefigge. Questi limiti della valutazione della qualità dei dati di addestramento possono avere come risultato dataset afflitti da rumore, pieni di errori e incoerenti, il che a sua volta può portare a comportamenti imprevisti a livello di sistema. Se non sono tenuti sotto controllo, gli errori e le incertezze relativi ai dati continuano ad accumularsi e si propagano fra i vari elementi di un sistema IA.

1.2.4 Debito tecnico

Nello sviluppo del software, il termine "debito tecnico" è una metafora per indicare problemi e costi del software sul lungo termine, derivanti dall'abbandono delle *best practices* (o buone pratiche), nella fase di sviluppo, a favore di soluzioni più facili e veloci. Le *best practices*

implementate comunemente nello sviluppo moderno del software (come il controllo delle versioni e i test di unità e sistema) non si trasferiscono facilmente al campo dell'IA, per la mancanza di procedure standard e per la difficoltà di definire test robusti per i modelli IA (Sculley et al., 2015).

La mancanza di strumenti comunemente accettati per il controllo delle versioni e per i test nel campo dell'IA in generale si traduce in un'adozione "a macchia di leopardo" delle pratiche di *Continuous Integration/Continuous Delivery* (CI/CD). Pipeline CI/CD affidabili, per esempio, richiedono un controllo esteso delle versioni. Non potendo controllare in modo affidabile le versioni di un sistema IA si possono generare problemi di robustezza, ed è il motivo per cui questa è una delle cause del problema della predicibilità nella sua definizione massima. Inoltre, il numero elevato e le interrelazioni delle componenti di un sistema IA rendono difficile controllarne i confini di astrazione. Questi aspetti diventano più problematici se modelli ML sono online e continuano ad apprendere durante l'uso. Nella misura in cui ostacola l'affidabilità e la tracciabilità dei sistemi IA, il debito tecnico inibisce la capacità degli osservatori di prevedere gli esiti di un sistema.

Dopo aver delineato qui gli aspetti fondamentali del problema della predicibilità, nel resto del libro, e in particolare nei capitoli 4-8, mi concentrerò sulle implicazioni etiche dell'uso dell'IA nella difesa. I prossimi due paragrafi di questo capitolo sono dedicati alla metodologia e all'ambito dell'analisi.

1.3 LA METODOLOGIA DEI LIVELLI DI ASTRAZIONE

L’analisi proposta in questo libro utilizza la metodologia dei LdA (Floridi, 2008). I LdA sono impiegati nell’ingegneria dei sistemi e nell’informatica per disegnare modelli di un dato sistema (Hoare, 1972; D. Heath, Allum, Dunckley, 1994; Diller, 1994; Jacky, 1997; Boca, 2014). Sono ampiamente utilizzati anche nell’etica digitale (Floridi, 2008; Floridi, Taddeo, 2016) e sono stati applicati per affrontare vari problemi fondamentali, per esempio per l’identificazione delle responsabilità dei fornitori di servizi online (Taddeo, Floridi, 2015), come guida nell’applicazione delle tecnologie di tracciamento durante la pandemia da Covid-19 (Morley, Cowls et al., 2020), per analizzare le possibilità della deterrenza nel cyberspazio (Taddeo, 2018c) e per considerare le implicazioni etiche della fiducia nelle tecnologie digitali (Taddeo, 2017b).

Il metodo si basa sul presupposto che qualsiasi sistema possa essere osservato concentrandosi su determinate proprietà specifiche e trascurandone altre. La scelta di quelle proprietà (gli “osservabili”) dipende dagli scopi dell’osservatore. Per esempio, per un tecnico interessato a perfezionare l’aerodinamica di un’automobile, gli osservabili possono essere la forma delle sue parti, il loro peso, la tipologia dei materiali utilizzati. Per un cliente interessato all’estetica della stessa automobile, gli osservabili possono essere invece il suo colore, gli interni, l’aspetto complessivo. Il tecnico e il cliente osservano la stessa auto (lo stesso sistema) a differenti LdA, che consentono loro di definire modelli differenti di quell’auto.

Un LdA è definito quindi come un insieme, finito ma non vuoto, di osservabili, accompagnato da una dichiarazione di quale sia la caratteristica precisa del sistema che il LdA rappresenta. Un LdA non riduce un’auto all’aerodinamica delle sue parti o al suo aspetto complessivo; è uno strumento che contribuisce a rendere esplicita la prospettiva dell’osservatore del sistema e a vincolarla solo agli elementi che sono funzionali per quella particolare osservazione e quel particolare scopo. Un LdA può avere un livello di granularità maggiore o minore: “La quantità di informazione in un modello varia con il LdA: un LdA di maggiore risoluzione o di granularità più fine produce un modello che

contiene più informazioni di un modello prodotto a un LdA superiore, o più astratto" (Floridi, 2008, p. 315).

Quando si considerano le sfide etiche dell'IA utilizzata nella difesa, ci si può concentrare su LdA diversi. Per esempio, si può decidere di prendere in considerazione solo i problemi etici che emergono durante la fase di progettazione e trascurare invece le fasi di sviluppo e d'uso del ciclo di vita dell'IA. Analogamente, le analisi etiche possono concentrarsi solo sull'intenzione d'uso o sugli effetti dell'uso dell'IA in questo campo. La scelta del LdA è orientata allo scopo: non esiste LdA corretto o non corretto *di per sé*, ma solo un LdA corretto o no secondo l'obiettivo dell'osservatore.

Visto l'obiettivo di questo libro, adotterò due LdA: LdA_{scopo} e LdA_{etica}. Gli osservabili di LdA_{scopo} sono gli *scopi immediati* dell'uso dell'IA. Gli osservabili di LdA_{etica} sono, per un dato scopo d'uso immediato, gli aspetti della progettazione, dello sviluppo e dell'uso dell'IA che possono portare a conseguenze etiche o non etiche. Vale la pena di sottolineare che lo scopo dell'uso non è la *funzione* di un sistema IA, dal momento che un sistema con la stessa funzione può porre problemi etici diversi quando usato per scopi diversi. Per esempio, un sistema di riconoscimento di immagini pone problemi diversi quando è utilizzato per il monitoraggio delle forme di vita marine e quando è implementato su un drone armato. La scelta di concentrarci qui sulle finalità d'uso anziché sulla funzione della tecnologia si basa su due motivi: la malleabilità delle tecnologie digitali e l'obiettivo generale del libro.

"Malleabilità" si riferisce al fatto che le tecnologie digitali, anche le più sofisticate, possono essere facilmente adibite a finalità differenti da quelle considerate in fase di progettazione (Moor, 1985, p. 269). Per la loro malleabilità, le sfide etiche delle tecnologie digitali (dell'IA in particolare) non sono definite tanto dalla funzione progettuale quanto dallo scopo per cui sono usate. Nel campo della difesa, questi scopi possono essere identificati chiaramente ed è probabile che diano forma sia agli usi correnti sia a quelli futuri dell'IA. Per questo la focalizzazione sullo scopo è più appropriata per sviluppare un'etica dell'IA nella difesa.

Per quanto riguarda il libro, il suo obiettivo non è definire una tassonomia onnicomprensiva delle tecnologie IA e delle relative implicazioni etiche, bensì offrire dei criteri per identificare i problemi etici collegati all'uso dell'IA nel campo della difesa, analizzarli e fornire

soluzioni e una guida efficace per affrontarli. Ecco perché la focalizzazione sullo scopo invece che sulla funzione è più appropriata nel nostro caso.

Utilizzando questi LdA, possiamo ora definire l'ambito dell'analisi etica dell'IA nella difesa. In questo libro esamineremo tre scopi d'uso dell'IA nella difesa: di sostegno e supporto; conflittuali e non cinetici; conflittuali e cinetici (Figura 1.1).

Figura 1.1 I tre scopi d'uso dell'IA per la difesa nazionale esaminati in questo libro. Lo schema è adattato da Tadde et al., 2021, p. 1710.

1.4 PROBLEMI ETICI DELL'USO DELL'IA PER LA DIFESA

I tre scopi d'uso dell'IA nel campo della difesa presentati sopra pongono problemi etici la cui difficoltà è progressivamente crescente, passando dagli usi di sostegno e supporto a quelli conflittuali e cinetici (vedi la Figura 1.1). Questo perché, insieme ai problemi etici relativi all'uso dell'IA (per esempio, trasparenza ed equità), dobbiamo considerare anche quelli relativi agli usi coercitivi (cinetici e non cinetici) di questa tecnologia e il loro impatto dirompente e distruttivo.

Come si vede nella Figura 1.2, ogni categoria d'uso presenta rischi etici specifici, ma eredita anche quelli delle categorie poste alla sua sinistra. Per esempio, gli usi conflittuali e non cinetici dell'IA pongono rischi per i diritti individuali, oltre a quello dell'escalation. Gli usi conflittuali e cinetici dell'IA pongono rischi per la trasparenza e l'autonomia umana, che appaiono già nella categoria "di sostegno e supporto", accanto a quelli per la protezione dei diritti, ma introducono anche quelli relativi al rispetto dei principi della Teoria della Guerra Giusta, della virtù militare, della dignità umana e della stabilità internazionale. Nei paragrafi che seguono esamineremo più in dettaglio i rischi etici fondamentali di ciascuna finalità d'uso.

Figura 1.2 Una mappa dei rischi etici relativi a ciascuna categoria d'uso dell'IA nella difesa. Lo schema è adattato da Taddeo et al., 2021, p. 1712.

1.4.1 Usi dell'IA per sostegno e supporto

L'uso di sistemi IA per scopi di difesa non conflittuali va dalle applicazioni nella cybersicurezza (US Army, 2017), in cui l'IA svolge un ruolo crescente nel garantire la robustezza e la resilienza dei sistemi, ai droni basati su IA per la ricognizione video, ai tag RFID (identificazione a radiofrequenza) sulle forniture alimentari (Lysaght, Harris, Kelly, 1988; Fraga-Lamas et al., 2016; Schubert et al., 2018) e all'uso dell'IA per l'analisi dell'intelligence (Blanchard, Taddeo, 2023).

Anche quando questo uso dell'IA non pone rischi per i diritti individuali e non è legato all'uso della forza, pone comunque sfide etiche importanti. Consideriamo l'uso dell'IA per rafforzare la robustezza dei sistemi. L'IA può aiutare nella verifica e validazione del software, liberando gli esperti umani da compiti noiosi e offrendo una maggiore velocità e accuratezza nei test dei sistemi (King et al., 2019). Per esempio, modelli generativi come i modelli linguistici di grandi dimensioni (LLM) hanno mostrato un grande potenziale per l'identificazione delle vulnerabilità.[5] In questo senso, l'IA può portare i test del software a un nuovo livello e rendere i sistemi più robusti. Bisogna però stare attenti al modo in cui utilizziamo l'IA in tale contesto, perché delegare i test all'IA può portare a una

completa dequalificazione del personale della difesa incaricato della verifica e validazione di sistemi e reti, ottenendo come esito finale una mancanza di controllo di questa tecnologia.

L'IA viene impiegata sempre di più anche per il rilevamento di minacce e anomalie nei sistemi informatici. Per esempio, nell'aprile 2017 la società di software DarkTrace ha lanciato Antigena, che utilizza il *machine learning* per individuare comportamenti anomali su una rete IT, bloccare le comunicazioni a quella parte del sistema e inviare un allarme. Questi servizi analizzano malware e virus, e alcuni possono mettere in quarantena le minacce e parti del sistema per indagini ulteriori. In certi casi, gli scanner delle minacce hanno accesso a file, email, dispositivi mobili e fissi, o addirittura ai dati del traffico su una rete. L'IA può essere utilizzata anche per l'autenticazione degli utenti mediante monitoraggio del comportamento e generazione di profili biometrici basati, per esempio, sul modo peculiare in cui un utente sposta il mouse ("BehavioSec", 2019). A volte questo può comportare il tracciamento di "dati dei sensori e delle interazioni fra umani e dispositivi dalla vostra app o dal vostro sito web. Viene registrato ogni tocco, ogni movimento del dispositivo o ogni gesto del mouse".[6]

Il rischio è chiaro. L'IA può migliorare la resilienza del sistema agli attacchi, ma questo richiede un monitoraggio molto esteso del sistema e dei suoi utenti e una raccolta di grandi quantità di dati per addestrare il modello in questione. Ciò può violare la privacy degli utenti, esporli a rischi ulteriori qualora la confidenzialità dei dati venga violata e creare un effetto di sorveglianza di massa (Taddeo, 2013, 2014b). Gli usi di sostegno e supporto dell'IA nella difesa e nella sicurezza creano difficoltà etiche simili a quelle individuate in altri campi, per esempio per quanto riguarda le violazioni della privacy. È importante però notare che questi problemi devono essere affrontati in specifico riferimento all'ambito della difesa. In questo caso, potrebbe rendersi necessario un bilanciamento fra interessi di Stato, uso della forza, difesa nazionale e rispetto dei diritti individuali.

L'IA viene utilizzata anche per migliorare la *situational awareness* e l'analisi dell'intelligence (di cui parleremo nel capitolo 3). Una *situational awareness* tempestiva è fondamentale per rafforzare la preparazione e prevenire minacce. Tuttavia, migliorare tale consapevolezza tramite l'IA può sollevare problematiche etiche, soprattutto considerando la natura ibrida delle minacce e le numerose

variabili in gioco. Le minacce, che possono appunto essere di natura ibrida, potrebbero coincidere con cambiamenti nelle circostanze politiche, economiche, strategiche, culturali e sociali che influenzano il difensore, e gli attacchi potrebbero essere lanciati da attori che operano con alleati, interessi, risorse e metodi mutevoli. Ciò richiede capacità di analisi in tempo reale e rilevamento delle anomalie sempre attive. L'IA offre molto a tal fine, dal momento che consente di esaminare grandi volumi di dati, ma la sfida principale consiste nel garantire che la raccolta e l'analisi su larga scala dei dati siano bilanciate da regolamenti e valori etici fondamentali, per evitare di minare la fiducia dei civili nelle istituzioni della difesa e della sicurezza, per esempio attraverso una sorveglianza eccessiva e ingiustificata o sistemi discriminatori.

L'analisi di grandi volumi di dati consente all'IA di estrarre informazioni a supporto della logistica, dei processi decisionali, nonché di *foresight analyses* (analisi prospettiche) e della governance interna. Questi usi dell'IA facilitano una gestione tempestiva ed efficace delle risorse umane e fisiche, migliorano la valutazione del rischio e supportano le decisioni. Per esempio, è stato riportato che gli ufficiali militari potrebbero avere solo tra gli otto e i dieci minuti per decidere se il lancio di un missile rappresenti una minaccia, condividere i risultati con gli alleati e valutare come rispondere.[7]

L'IA sarebbe di grande aiuto in un simile scenario, perché potrebbe integrare in tempo reale i dati provenienti da satelliti e sensori ed elaborare informazioni chiave che potrebbero contribuire al processo decisionale. Tuttavia, al grande potenziale dell'IA a scopi di supporto si affiancano seri rischi etici, che comprendono bias se i sistemi IA sono parziali, mancanza di trasparenza e responsabilità, bias da automazione, nonché *responsibility gap* (vuoto di responsabilità), sollevando interrogativi sul modo in cui questi sistemi sono progettati e integrati nei processi decisionali.

1.4.2 Usi conflittuali e non cinetici dell'IA

Con l'escalation delle minacce contro infrastrutture pubbliche e confini nazionali, cresce anche la necessità di strategie di difesa in grado di affrontarle. Questo vale anche per le minacce cibernetiche. Il Regno Unito e gli Stati Uniti hanno utilizzato metodi di difesa cibernetica *attiva* (o

defend forward) per consentire agli esperti informatici di neutralizzare o distrarre i virus con falsi bersagli e di penetrare nei sistemi degli hacker per cancellare dati o per distruggerli completamente. Nel 2016, il Regno Unito ha annunciato un investimento di 1,9 miliardi di sterline e un piano quinquennale per combattere le minacce cibernetiche, investimento aumentato a 2,6 miliardi nel 2022.[8] Nel 2020, il Regno Unito ha poi costituito la National Cyber Force, iniziativa congiunta fra ministero della Difesa e Government Communications Headquarters, con il compito di individuare attori stranieri ostili. Su scala internazionale, ora la NATO può contare su cybereffetti sovrani (cioè cyberattacchi lanciati dagli Stati membri) in risposta a cyber-attacchi, come concordato durante il summit di Bruxelles,[9] per consentire all'alleanza di punire cyberattacchi (una volta che questi siano attribuiti con chiarezza) ed evitare che gli attaccanti colpiscano ancora.

L'IA continuerà a rivoluzionare queste attività. I cyberattacchi e le risposte diventeranno più veloci, più precisi e più distruttivi. L'IA amplierà le capacità di individuazione dei bersagli da parte degli attaccanti, consentendo loro di utilizzare dati più complessi e ricchi per scegliere i bersagli e progettare gli attacchi, e l'IA nel malware può modificare la natura e la conduzione di un attacco. Esistono già prototipi di cyberarmi abilitate dall'IA, fra cui malware autonomi per corrompere immagini mediche e colpire veicoli autonomi (Mirsky et al., 2019; Zhuge et al., 2007). Per esempio, IBM ha creato un prototipo di malware autonomo, DeepLocker, che utilizza una rete neurale per selezionare i suoi bersagli e mimetizzarsi fino a che non raggiunge la destinazione ("DeepLocker", 2018). Sistemi di cybersicurezza autonomi e semi-autonomi dotati di un "manuale" di risposte predeterminate a un'attività, che vincolano l'agente ad azioni note, sono disponibili sul mercato ormai da qualche anno ("DarkLight offers first of its kind artificial intelligence to enhance cybersecurity defenses", 2017). Sono in commercio anche sistemi autonomi in grado di apprendere i comportamenti degli avversari e di generare falsi bersagli e "specchietti per le allodole" per sviare i portatori di minacce ("Acalvio autonomous deception", 2019). Dalle ricerche è emerso anche che gli LLM possono sfruttare autonomamente le vulnerabilità *zero-day*[10] del mondo reale, con conseguenze gravi per la cybersicurezza (Fang et al., 2024).

Se gli Stati acquisiscono e sviluppano capacità di IA per la difesa nazionale, lo stesso vale per i loro avversari (Taddeo, Floridi, 2018a). Come vedremo nel capitolo 4, questo può portare a un paradosso della sicurezza: gli Stati che aumentano le proprie capacità IA possono essere percepiti come una minaccia da altri; da questo potrebbero derivare tensioni internazionali, che possono provocare un'intensificazione di cyberattacchi e risposte, con rischi di escalation in grado di portare a conseguenze cinetiche (Taddeo, 2018c). Per mitigare questi rischi, è fondamentale che gli usi dell'IA rispettino i principi chiave della Teoria della Guerra Giusta, che sta alla base delle norme internazionali – come lo Statuto delle Nazioni Unite,[11] le Convenzioni dell'Aia e di Ginevra[12] e il Diritto internazionale umanitario[13] – e fissa i parametri dei dibattiti etici e politici sui conflitti cibernetici. È fondamentale che l'uso dell'IA per scopi conflittuali e non cinetici rispetti i principi della proporzionalità della risposta, discrimini fra bersagli legittimi e no, garantisca qualche forma di riparazione quando vengono commessi errori (Taddeo, 2012a, 2012b, 2014a) e mantenga la responsabilità e il controllo nella catena di comando. Le considerazioni etiche sull'uso conflittuale e non cinetico dell'IA devono contribuire alla comprensione di come applicare la Teoria della Guerra Giusta nel cyberspazio, e devono essere utilizzate per orientare la discussione sulla regolamentazione del comportamento degli Stati in questo nuovo dominio della difesa (Taddeo, 2016a, 2017a). Questi aspetti saranno ulteriormente discussi nei capitoli 4 e 5.

1.4.3 Usi conflittuali e cinetici dell'IA

Nel considerare le implicazioni etiche degli usi conflittuali e cinetici dell'IA per la difesa, tendenzialmente l'analisi si è concentrata sulla combinazione dell'IA con strumenti in grado di causare danni letali agli esseri umani e di distruggere oggetti fisici in modo completamente autonomo. Gli usi dell'IA per gli scopi che consideriamo in questo paragrafo, però, variano molto: si va dall'automazione di varie funzioni di un sistema d'arma a sistemi che seguono le istruzioni pre-programmate da un essere umano, a sistemi d'arma totalmente autonomi (*autonomous weapon systems*, AWS) che possono identificare, selezionare e colpire bersagli senza input umano diretto. Un esempio è un sistema sviluppato per la Royal Navy del Regno Unito e chiamato STARTLE,[14] che supporta i

processi decisionali umani con software di *situational awareness* e che monitora e valuta potenziali minacce con una combinazione di tecniche IA. Analogamente, l'Advanced Targeting & Lethality Automated System[15] sviluppato per l'esercito degli Stati Uniti supporta gli operatori umani nell'identificazione delle minacce e nello stabilire la priorità di potenziali bersagli. A quanto è stato riferito, Israele ha sviluppato e utilizzato The Gospel e Lavender, due sistemi IA per l'identificazione di bersagli umani.[16] Questi usi dell'IA pongono seri interrogativi e problemi etici, la cui gravità varia a seconda del grado di autonomia del sistema IA, delle sue funzionalità di apprendimento, dello scopo d'uso e del livello di controllo umano.

Anche in questo caso, il problema centrale è garantire che questi usi dell'IA rispettino i principi della Teoria della Guerra Giusta, per esempio quelli di necessità, proporzionalità e distinzione. I problemi riguardano sia le modalità di impiego sia le capacità dei sistemi usati. Dal punto di vista procedurale, i sistemi IA devono essere impiegati seguendo procedure che garantiscano un'appropriata supervisione umana e la possibilità di intervenire e cambiare le decisioni dell'IA. Tecnicamente, per esempio, i sistemi IA devono essere in grado di distinguere fra combattenti e non combattenti e devono riconoscere i segnali di resa comunemente accettati nei conflitti armati. Questi sono criteri problematici da rispettare, dato che l'IA non è ancora in grado di analizzare il contesto fino a questo punto e, in alcune situazioni, la sua capacità di riconoscere bersagli legittimi è significativamente peggiore di quella degli esseri umani (Sharkey, 2010, 2012a, 2012b; Tamburrini, 2016). Sono d'accordo. Altrove però ho sostenuto che lo stato dello sviluppo dell'IA è sostanzialmente irrilevante per la soluzione di questi problemi; i principi della Teoria della Guerra Giusta saranno sempre a rischio quando si considerano gli usi conflittuali e cinetici, a causa del problema intrinseco della predicibilità dell'IA e della conseguente limitazione del controllo umano su questa tecnologia (Blanchard, Taddeo, 2022a, 2022b). Anche se le tecnologie dell'IA progredissero al punto da interpretare correttamente i contesti, la limitata predicibilità dei loro esiti introduce una fonte (non nulla) di rischio che può essere considerata inaccettabile quando si considerano gli usi cinetici. Per questo il dibattito sull'ammissibilità delle armi autonome deve concentrarsi su soglie di rischio accettabili, più che sullo stato di sviluppo delle tecnologie IA. Torneremo su questo punto nel capitolo 8.

Le difficoltà nell'attribuire la responsabilità morale per le azioni di un sistema IA è un altro problema etico fondamentale. È problematico per tutte le tre categorie di usi dell'IA, ma è particolarmente preoccupante nel caso di quelli conflittuali e cinetici, visti i rischi (Sparrow, 2007). Diventa un problema ancora più pressante quando si considera il rispetto dovuto in guerra agli avversari e alla loro dignità. Trattare rispettosamente gli avversari è un modo importante per mantenere la moralità della guerra (Nagel, 1972) e la relazione interpersonale con l'avversario è fondamentale a questo scopo. Se l'uso degli AWS tronca tale relazione, dobbiamo chiederci se questi sistemi minino la dignità delle persone che prendono a bersaglio (e forse anche di quelle che li usano) e se ciò porti a una forma di omicidio moralmente problematico (Asaro, 2012; Docherty, 2014; Sharkey, 2019; Johnson, Axinn, 2013; Sparrow, 2016; O'Connell, 2014; Ekelhof, 2019).

I problemi etici riguardano anche l'impatto degli AWS sulla stabilità internazionale. Da un lato, possono ridurre l'intervallo temporale delle ostilità tra Stati, e quindi favorire la stabilità; possono anche essere un deterrente efficace per i possibili avversari. D'altro lato, gli AWS possono portare a una guerra ingiusta e promuovere l'instabilità internazionale, per esempio se l'uso asimmetrico fa sì che la parte più debole ricorra all'insurrezione e a tattiche terroristiche (Sharkey, 2012a, 2012b). Poiché in generale il terrorismo è considerato una forma di guerra ingiusta (o, peggio, un atto di omicidio indiscriminato), l'uso degli AWS può portare a una maggiore incidenza della violenza immorale. Tornerò su questi punti nei capitoli 6, 7 e 8.

1.5 CONCLUSIONE

Ho già detto che l'obiettivo del libro è fornire un'analisi etica sistematica degli usi dell'IA nella difesa. Per onestà nei confronti di chi legge, devo dire che c'è un altro obiettivo, seppure secondario: promuovere un approccio proattivo dell'establishment della difesa nei confronti dell'etica dell'IA nella difesa.

La difesa, cinetica o no, è diventata digitale. I dati e le tecnologie per raccogliere, analizzare e comunicare (e i sistemi IA che sempre più sostengono e producono dati) sono diventati asset fondamentali che generano cambiamenti profondi nelle dottrine e nelle strategie militari. La rivoluzione digitale sta trasformando la difesa così come ha trasformato la nostra vita quotidiana e le nostre società. Questi cambiamenti sono dirompenti e, se non sono compresi bene, possono portare a conseguenze inattese e indesiderate e, soprattutto, a violazioni dei valori e dei diritti fondamentali delle nostre società.

Il caso dei conflitti cibernetici è emblematico. Il vuoto normativo per il comportamento degli Stati nel cyberspazio è il risultato di molte cause, fra le quali è fondamentale la percezione dell'etica, nel migliore dei casi, come un'appendice alla discussione sulla trasformazione digitale della difesa e, nel peggiore, come un ostacolo irritante alla necessità di ottenere un vantaggio sull'avversario. A oltre un decennio dal summit NATO del 2014, questo approccio ha favorito comportamenti aggressivi nel cyberspazio, dove gli attacchi gestiti e sponsorizzati dagli Stati diventano sempre più numerosi, con attacchi come SolarWinds, e come quello contro Colonial Pipeline e quello all'Ufficio di gestione del personale degli Stati Uniti,[17] che hanno causato ampi danni alle infrastrutture civili, violando diritti fondamentali e aggravando tensioni geopolitiche (Taddeo, 2017a).

L'adozione crescente dell'IA nella difesa nelle tre categorie d'uso descritte all'inizio di questo capitolo rende la necessità di regolarne gli impieghi sempre più urgente. Una necessità riconosciuta da diverse istituzioni della difesa – per esempio i ministeri della Difesa inglese, francese e americano, e anche la NATO – che hanno già definito principi etici e commissioni per l'etica dell'IA. Sono sforzi iniziali che richiedono molti altri passi per essere completati con successo.

Per essere efficaci questi sforzi devono fondarsi su una comprensione adeguata dei cambiamenti concettuali prodotti dalla rivoluzione digitale e delle loro implicazioni etiche. È un tema ben noto nell'etica digitale. Come dice Moor: "Anche se un problema […] inizialmente può sembrare chiaro, un po' di riflessione mette in luce un *pasticcio concettuale*. Ciò che serve in tali casi è un'analisi che offra un *quadro concettuale coerente entro il quale formulare una norma per l'azione*" (1985, p. 266, corsivo mio).

L'obiettivo del libro è fornire questo quadro e con esso mostrare che l'etica dell'IA nella difesa non è una mera appendice o un ostacolo per lo sfruttamento del potenziale di questa tecnologia. È, invece, una risorsa imprescindibile da coltivare per le organizzazioni di difesa delle democrazie liberali, per far sì che la loro difesa non passi per la violazione dei nostri valori fondamentali.

---

1. "'Lavender': The AI machine directing Israel's bombing spree in Gaza", in *+972 Magazine*, 3 aprile 2024, https://www.972mag.com/lavender-ai-israeli-army-gaza/ (ultimo accesso 3 agosto 2024).
2. S. Bendett, "Roles and implications of AI in the Russian-Ukrainian conflict", Center for a New American Security, 20 luglio 2023, https://www.cnas.org/publications/commentary/roles-and-implications-of-ai-in-the-russian-ukrainian-conflict (ultimo accesso 3 agosto 2024).
3. Vale la pena di notare che ogni applicazione di IA avrà modelli di dati diversi o differenti approcci ML (dall'apprendimento con supervisione all'apprendimento per rinforzo), o sensori diversi, e questo influirà sul livello di automazione incorporato nel sistema IA. A sua volta, il livello di automazione nel sistema influirà sulla capacità dell'agente umano di vagliare e comprendere i suoi output.
4. "A data-centric view of technical debt in AI", Data-Centric AI, https://datacentricai.org/data-in-deployment/ (ultimo accesso 3 agosto 2024).
5. "Security implications of ChatGPT", Cloud Security Alliance, 8 febbraio 2023, https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt/ (ultimo accesso 2 agosto 2024).
6. "Denuvo Unbotify", ir.deto, http://www.unbotify.com (ultimo accesso 5 giugno 2024).
7. "The next major defense challenge", KPMG, 11 giugno 2018, https://kpmg.com/ph/en/home/insights/2018/06/the-next-major-defense-challenge.html (ultimo accesso 3 agosto 2024).
8. "Government cyber security strategy", UK Government, https:/assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1049825/government-cyber-security-strategy.pdf (ultimo accesso 3 agosto 2024).
9. "NATO's role in cyberspace", 12 febbraio 2024, https://www.nato.int/docu/review/articles/2019/02/12/natos-role-in-cyberspace/index.html (ultimo accesso 3 agosto 2024).

10. Falle di sicurezza in software o sistemi informatici non ancora note al produttore e per le quali i rimedi non sono ancora disponibili.

11. UN Charter, https://www.un.org/en/sections/un-charter/un-charter-full-text/ (ultimo accesso 3 agosto 2024).

12. Collection Military Legal Resources, Library of Congress, https://www.loc.gov/rr/frd/Military_Law/pdf/ASubjScd-27-1_1975.pdf (ultimo accesso 3 agosto 2024).

13. International Committee of the Red Cross, https://www.icrc.org/en/doc/resources/documents/misc/57jm93.htm (ultimo accesso 3 agosto 2024).

14. "Complex information fusion and advanced threat warning system", in *Roke*, https://www.roke.co.uk/products/startle (ultimo accesso 3 agosto 2024).

15. "ATLAS: Killer robot? No. Virtual crewman? Yes", in *Breaking Defense*, 4 marzo 2019, https://breakingdefense.com/2019/03/atlas-killer-robot-no-virtual-crewman-yes/ (ultimo accesso 30 marzo 2025).

16. "'Lavender': The AI machine directing Israel's bombing spress in Gaza", in *+972 Magazine*, 3 aprile 2024, https://www.972mag.com/lavender-ai-israeli-army-gaza/ (ultimo accesso 3 agosto 2024).

17. https://www.reuters.com/world/us/data-237000-us-government-employees-breached-2023-05-12/#:~:text=Two%20breaches%20at%20the%20U.S.,5.6%20million%20of%20those%20individuals.

2

PRINCIPI ETICI PER L'IA NELLA DIFESA

2.1 INTRODUZIONE

Dopo aver presentato la metodologia e l'ambito dell'analisi proposta in questo libro, esaminerò ora gli approcci esistenti all'etica dell'IA nella difesa. L'impatto dell'IA su tutte le funzioni della difesa nazionale ormai è chiaro e gli interrogativi su come promuovere l'uso etico delle tecnologie IA in questo campo sono diventati ancora più urgenti. Tutto ciò ha indotto varie istituzioni della difesa a considerare i rischi etici e a definire misure per mitigarli. Gli sforzi in questo settore sono ancora agli inizi, e fino al 2023 solo tre istituzioni – il Department of Defence (DoD) degli Stati Uniti (Defense Innovation Board, 2019), il Ministry of Defence (MoD) del Regno Unito (Ministry of Defence, 2022) e la NATO[1] – hanno adottato ufficialmente principi etici per l'IA (vedi la Tabella 2.1). I tre gruppi di principi presentano una coerenza notevole, sia per il loro contenuto sia per i loro limiti. Due dei limiti sono di particolare rilevanza per questo capitolo.

Tabella 2.1 Un riepilogo dei principi per l'uso etico dell'IA nella difesa proposti dal DoD degli USA, dal MoD del Regno Unito e dalla NATO.

Principi	**US DoD**	**UK MoD**	**NATO**
	Responsabile	Centrata sugli esseri umani	Legalità
	Equa	Responsabilità	Responsabilità e imputabilità
	Tracciabile	Comprensione	Esplicabilità e tracciabilità
	Reliable	Mitigazione dei bias e dei danni	Governabilità
	Governabile	*Reliable*	Mitigazione dei bias

Il primo dipende dall'approccio che sta alla base di questi principi: ossia l'IA responsabile (RAI), che mira a promuovere la progettazione, lo sviluppo e l'uso responsabili dei sistemi IA, cioè un ciclo di vita responsabile. Tornerò su questo punto nel paragrafo 2.3; qui mi limito a dire che questa impostazione non è la più adatta per l'uso dell'IA in ambiti ad alto rischio. Si basa sull'approccio *responsible research and innovation* (RRI; vedi, per esempio, Stilgoe, Owen, Macnaghten, 2013), che intende promuovere la riflessione critica sulle implicazioni potenziali della ricerca, ma non offre una guida concreta rispetto a ciò che si deve, o non si deve, fare per mitigare i rischi etici in casi specifici. Tuttavia, una guida concreta è *esattamente* quello che serve per promuovere l'uso etico dell'IA nella difesa.

Il secondo limite riguarda ciò che i principi prescrivono, che a volte sembra tanto ovvio da apparire ridondante. Per esempio, uno dei principi della NATO è che l'uso dell'IA deve essere "legale"; i principi degli USA sottolineano che l'IA deve essere utilizzata in modo da risultare "governabile"; i principi del Regno Unito stabiliscono che l'uso dell'IA deve essere "centrato sugli esseri umani". Non c'è nulla di controverso nell'idea che i sistemi IA debbano essere legali, governabili e centrati sugli esseri umani: anzi, viene da chiedersi quali mai potrebbero essere le alternative. Qui, la scelta di concentrarsi su punti ovvi penalizza la possibilità di fornire principi etici per l'IA che siano più significativi, focalizzati chiaramente sui rischi che l'IA pone per i valori e i diritti delle società democratiche e su principi dipendenti dal dominio, in particolare quelli articolati nella Teoria della Guerra Giusta. Il presente capitolo ha l'obiettivo di definire un insieme di principi alternativi, centrati su questi valori e diritti, e di offrire una metodologia che permetta di tradurre questi principi in prassi.

Il secondo limite dei principi è conseguenza anche del livello molto alto a cui sono formulati. Qualcuno ha sostenuto che l'alto livello dei principi sia un vizio fondamentale, che li rende troppo astratti per guidare effettivamente la progettazione, lo sviluppo e l'uso dei sistemi IA nella pratica (Coldicutt, Miller, 2019; Peters, 2019) o per informare i processi decisionali. Per parte mia, non vedo il livello alto dei principi etici per l'IA come uno svantaggio: sono principi di carattere fondazionale, non linee guida; offrono un inquadramento dei valori in gioco dipendente dal dominio. Sono perciò essenziali per orientare gli sforzi tesi a garantire usi

eticamente corretti dell'IA, ma da soli non sono sufficienti. In altre parole, questi principi etici offrono una bussola, non una mappa. Le critiche che ne mettono in dubbio l'efficacia per via del loro alto livello sono sbagliate, perché intendono quei principi come necessari e sufficienti per definire un ciclo di vita eticamente corretto per l'IA. I principi etici, invece, sono parte di sforzi più sistematici che sono necessari per promuovere un ciclo di vita eticamente corretto per l'IA in ambiti ad alto rischio.

Per essere efficaci, i principi etici dell'IA devono essere accompagnati da metodologie appropriate, che offrano una guida dominio-specifica su come interpretarli e applicarli (Taddeo, Floridi, 2018b), per garantire che ogni fase del ciclo di vita dell'IA rispetti i principi etici dell'organizzazione che ne sviluppa e supervisiona l'uso. Riconoscendo questa necessità, la letteratura sull'etica dell'IA è passata dal *che cosa* al *come* (Floridi, 2019, p. 185), producendo un corpus crescente di studi focalizzati sullo sviluppo di strumenti e processi per implementare principi etici dell'IA.[2]

Concentrandosi direttamente su strumenti e soluzioni applicabili, però, la letteratura in materia lascia senza risposta questioni normative cruciali. Per esempio, quando vengono applicati a casi specifici, i principi etici dell'IA possono generare tensioni che richiedono compromessi non risolvibili con il ricorso ai soli principi o strumenti implementativi (Whittlestone et al., 2019). Pensiamo, per esempio, al caso in cui la trasparenza entra in conflitto con problemi di sicurezza nazionale. Analogamente, la scelta di uno strumento per l'applicazione dei principi etici dell'IA coinvolgerà decisioni normative, che non possono essere prese facendo riferimento solo ai principi (Blanchard, Thomas, Taddeo, 2023). Per esempio, la scelta di uno strumento di auditing etico richiede una decisione riguardo al tipo di metrica, qualitativa o quantitativa, da utilizzare per la valutazione. A sua volta il tipo di metrica determina l'ambito dell'audit. Ciò significa che occorre una guida sulle metriche da scegliere in situazioni specifiche. In breve, esiste un passaggio intermedio fra i principi etici dell'IA (ad alto livello) e gli strumenti di etica dell'IA; questo passaggio riguarda la specificazione di una metodologia per l'implementazione, vale a dire l'interpretazione di quei principi nella prassi. Fornire questa metodologia è il secondo obiettivo del capitolo.

Nel resto del capitolo analizzerò i principi etici dell'IA adottati dall'industria della difesa degli Stati Uniti, nel paragrafo 2.2, e la metodologia proposta per implementarli, nel paragrafo 2.3, per ricavarne alcuni insegnamenti preziosi, prima di offrire un nuovo insieme di principi per l'etica dell'IA nella difesa. Presento questi principi nel paragrafo 2.4. Nel paragrafo 2.5 introdurrò una metodologia per estrarre linee guida eticamente valide, applicabili ed efficaci per l'uso dell'IA nella difesa a partire dai principi etici proposti. Il paragrafo 2.6 trarrà le conclusioni del capitolo.

2.2 PRINCIPI ETICI PER L'USO DELL'IA

Fra i principi etici pubblicati, quelli del DoD degli Stati Uniti sono gli unici accompagnati da un documento (Defense Innovation Board, 2019) che ne espone le motivazioni e offre raccomandazioni per soluzioni di governance nell'implementazione dei principi (da qui in avanti vi farò riferimento semplicemente come al "documento di supporto"). Il DoD degli Stati Uniti è anche l'unica istituzione della difesa che abbia pubblicato linee guida ufficiali per l'operazionalizzazione dei principi (vedi il paragrafo 2.3).[3] Per questo mi concentrerò solo sui principi del DoD, lasciando da parte quelli del MoD del Regno Unito e della NATO, che sono simili per ambito e contenuti, ma sono meno dettagliati. Il DoD degli Stati Uniti offre cinque principi, che prescrivono che l'uso dell'IA nella difesa debba essere responsabile, equo, tracciabile, *reliable*[4] e governabile. I sottoparagrafi che seguono analizzano i singoli principi, considerandone gli aspetti positivi e i limiti.

2.2.1 Usi responsabili dell'IA

Il primo principio stabilisce che gli usi dell'IA devono essere responsabili:

> Gli esseri umani devono esercitare livelli appropriati di giudizio e rimanere responsabili dello sviluppo, della messa in campo, dell'uso e degli esiti dei sistemi IA del DoD. (Defense Innovation Board, 2019, p. 8)

Questo principio non è controverso ed è coerente con quelli presentati da vari quadri di riferimento etico (Department for Digital, Culture, Media & Sport, 2018, p. 5; Gavaghan et al., 2019, p. 41; Japanese Society for Artificial Intelligence, 2017, p. 3; Ministry of Defence, 2022). Nel documento di supporto, la raccomandazione su come implementare questo principio propone un sistema di responsabilità a tre livelli, dove il primo riguarda gli esseri umani che controllano

> la progettazione, la definizione dei requisiti, lo sviluppo, l'acquisizione, i test, la valutazione e l'addestramento per qualsiasi sistema del DoD, inclusi quelli di IA. (Defense Innovation Board, 2019, p. 27)

Il secondo livello di responsabilità riguarda l'uso dell'IA nei conflitti (cinetici o no), con responsabilità assegnate alla struttura di comando e controllo, nella misura in cui comandanti e operatori hanno "informazioni appropriate sul comportamento di un sistema, un addestramento pertinente, intelligence e *situational awareness*" (*ibidem*, p. 28). Il terzo livello di responsabilità è relativo ai meccanismi di riparazione dopo la conclusione delle ostilità. Questo livello riguarda sia il DoD sia il settore privato che fornisce la tecnologia IA per la difesa. Il documento di supporto specifica che la responsabilità umana in questo caso si basa su un "giudizio umano appropriato".

Questo approccio è solo parzialmente corretto. La definizione di giudizio "appropriato" è vaga e, perciò, problematica, specialmente se si considerano i problemi posti dalla mancanza di trasparenza e predicibilità di alcuni sistemi IA. Allo stesso tempo, l'attribuzione di responsabilità secondo il sistema a tre livelli rischia di scaricare le responsabilità al primo livello (sviluppo e addestramento), nella misura in cui le conseguenze non volute di sistemi IA possono essere ricondotte a problemi di progettazione e sviluppo. Questo può avere un effetto dannoso sul modo in cui gli attori coinvolti nel comando e controllo possono percepire le proprie responsabilità rispetto all'uso dell'IA.

2.2.2 Usi equi dell'IA

Il principio degli usi equi dell'IA stabilisce che:

> Il DoD deve prendere misure ponderate per evitare bias non voluti nello sviluppo e nell'uso di sistemi IA, per il combattimento o no, che provochino involontariamente danni a persone. (Defense Innovation Board, 2019, p. 8)

Il principio si concentra su questioni legate a equità e giustizia, ma evita di utilizzare direttamente questi termini. Nel documento di supporto, il motivo del mancato uso del termine "equità" è che:

> Questo principio deriva dal mantra del DoD che gli scontri non devono essere equi, perché il DoD mira a creare le condizioni per mantenere un vantaggio non equo su qualsiasi potenziale avversario, per aumentare la probabilità che questo funga da deterrente ed eviti lo scoppio di conflitti. (*Ibidem*, p. 31)

Il documento continua:

> Il DoD deve avere sistemi IA che siano appropriatamente orientati a colpire con maggiore successo certi combattenti avversari e minimizzino qualsiasi impatto dannoso su civili, non combattenti o altri individui che non devono costituire dei bersagli. (*Ibidem*, p. 33)

Questo scostamento dal modo in cui solitamente si intende il termine "equità" nei sistemi IA, quindi, è motivato dalla percezione della natura peculiare della difesa; ma è fuorviante, perché induce a pensare che la necessità di trovare un vantaggio sull'avversario possa giustificare usi non equi, o addirittura ingiusti, dell'IA. Non è così, invece, perché il principio dell'equità non stabilisce che l'esito del conflitto sia paritario – "a somma zero" – tra gli interessi delle parti contrapposte; si riferisce invece a esiti che riguardano i diritti e i doveri individuali, in uno scenario specifico. In tal senso si riferisce al principio di giustizia, secondo il quale gli individui devono essere trattati nella stessa maniera, a meno che siano diversi in modi rilevanti per la situazione in cui sono coinvolti (Rawls, 2005).

Il principio di giustizia vale anche per la guerra in cui, in effetti, distinguiamo fra condotta *giusta* e *ingiusta* nella difesa, e puniamo la seconda. Esistono differenze fra i modi in cui il principio di giustizia si applica in contesti civili e nei conflitti. Nella difesa, il principio di giustizia deve essere bilanciato con quello della necessità militare, che consente l'uso della forza (letale, nei limiti definiti dai principi di proporzionalità e distinzione), se è ritenuto necessario per uno scopo militare legittimo, come la sconfitta del nemico o la protezione delle proprie forze armate.

Nella misura in cui il principio del DoD si concentra solo sugli usi equi dell'IA e li caratterizza soltanto rispetto ai rischi di bias e discriminazione indebita nei confronti del personale del DoD, non considera lo spettro più ampio, più rilevante, dei rischi di usi ed esiti ingiusti dell'IA nella difesa, né promuove misure che li mitighino. Consideriamo, per esempio, l'uso dei LAWS: come vedremo nel capitolo 8, queste armi possono essere utilizzate per acquisire un vantaggio sull'avversario ma, dato il problema della predicibilità, il controllo sui loro effetti è limitato. L'uso dei LAWS pone rischi seri per il principio di distinzione, colpendo non combattenti. In base alla Teoria della Guerra Giusta e del diritto umanitario internazionale (IHL), questo sarebbe un uso dell'IA ingiusto e quindi non ammissibile. Nella misura in cui il principio del DoD si concentra solo sugli usi equi dell'IA non contempla questo rischio, perciò offre una guida

limitata per l'uso dell'IA nella difesa. I tentativi di stabilire gli usi etici dell'IA nella difesa devono definire principi per gli usi giusti dell'IA che sono rilevanti in questo campo e coerenti con i principi forniti dalla Teoria della Guerra Giusta (Taddeo, 2014a). In caso contrario, rischiano di diventare irrilevanti.

2.2.3 Tracciabilità

Il principio della tracciabilità affronta indirettamente i problemi etici posti dalla mancanza di trasparenza dell'IA. Afferma che:

> La disciplina dell'ingegneria IA del DoD deve essere avanzata al punto che gli esperti tecnici possiedano una comprensione appropriata della tecnologia, dei processi di sviluppo e dei metodi operativi dei suoi sistemi IA, ivi comprese metodologie, fonti di dati, procedure e documentazione di progettazione trasparenti e *auditable*. (Defense Innovation Board, 2019, p. 8)

È da notare che il principio non si focalizza sulla trasparenza della tecnologia *in sé*, ma sulla definizione di un livello minimo di competenze del personale del DoD, che deve avere una comprensione appropriata dei suoi sistemi IA, inclusa la tracciabilità dei processi e delle decisioni dei sistemi IA, sia nelle fasi di sviluppo sia in quelle operative. Come è specificato nel documento di supporto, la tracciabilità nella fase di sviluppo si riferisce alla raccolta e alla condivisione con gli stakeholder appropriati di "metodologia di progettazione, documenti di progetto rilevanti e fonti di dati" (*ibidem*, p. 34), mentre, nella fase operativa, comprende forme di monitoraggio, auditing e trasparenza dei processi. Come viene specificato nel documento di supporto:

> Alcuni sistemi possono richiedere non soltanto revisioni degli accessi degli utenti, ma anche la documentazione dell'uso e dei suoi scopi. Questo requisito può mitigare i danni legati all'uso *off-label* di un sistema IA, e anche rafforzare il principio di responsabilità. In breve, il DoD dovrà ripensare come traccia i suoi sistemi IA, chi ha accesso a particolari dataset e modelli, e se quegli individui li riutilizzano in altre aree applicative. (*Ibidem*, p. 35)

Senza trasparenza, la tracciabilità sarà molto limitata. Può promuovere usi responsabili dell'IA e in parte compensa la mancanza di trasparenza di questa tecnologia, ma non consente di ovviare ai rischi legati a questa mancanza. Pensiamo, per esempio, al problema della predicibilità. Se i sistemi IA sono scatole nere, è difficile prevedere i loro comportamenti in

contesti d'uso specifici e vagliare gli esiti quando questi sono non voluti o non attesi. È problematico che i documenti degli Stati Uniti si concentrino solo sulla tracciabilità, evitando di discutere la mancanza di trasparenza di molti sistemi IA, e non offrano alcun suggerimento rilevante per mitigarla, per esempio imponendo la valutazione della trasparenza di sistemi IA diversi e la scelta, fra i sistemi disponibili per un dato compito, di quello più trasparente.

Vale la pena di sottolineare anche che tracciabilità e trasparenza sono elementi *infraetici* (Floridi, 2017); in altre parole, il loro valore è determinato in rapporto al loro impatto sui principi etici e i diritti. Per questo la focalizzazione sulla tracciabilità o la trasparenza deve essere specificata rispetto all'esigenza di indagare i processi e all'attribuzione di responsabilità. Senza questo riferimento, non è chiaro quali informazioni debbano essere divulgate (e a chi) e quindi quale documentazione debba essere mantenuta. Formulate in termini generici, trasparenza e tracciabilità rischiano di favorire le cattive pratiche (Floridi, 2019), come lo "shopping etico" (cioè, combinare principi etici, linee guida, codici ecc. per giustificare comportamenti preesistenti); il *blue-washing* (cioè, fare affermazioni non comprovate o fuorvianti sul proprio impegno nell'etica dell'IA), o implementare solo superficialmente misure di etica dell'IA; il *lobbying* (cioè, fare leva sui principi dell'etica dell'IA per ritardare o evitare una buona e necessaria legislazione in materia); o l'*ethics shirking* (cioè, fare affermazioni non comprovate o fuorvianti sui valori etici, o implementare misure superficiali a favore dei valori etici) (*ibidem*, p. 186).

2.2.4 *Reliability* e governabilità

Il principio del DoD che riguarda la *reliability* dell'IA stabilisce che:

> I sistemi IA del DoD devono avere un dominio d'uso esplicito e ben definito e la sicurezza e la robustezza di tali sistemi devono essere testate e garantite lungo tutto il ciclo di vita in quel dominio d'uso. (Defense Innovation Board, 2019, p. 8)

Il documento di supporto sottolinea l'esigenza di sviluppare un'IA *reliable* (anziché degna di fiducia), le cui sicurezza e robustezza "devono essere testate e garantite" (*ibidem*). Questo principio è orientato specificamente a promuovere la verifica e la validazione dell'IA e a migliorarne la robustezza. Si tratta di un requisito fondamentale per l'uso

dell'IA nella difesa, che è importante citare esplicitamente per riaffermare la necessità di monitorare i sistemi IA (ne parleremo più approfonditamente nel paragrafo 2.4.5).

Allo stesso tempo, il documento di supporto evidenzia quanto è importante che agenti umani siano in grado di bloccare o disattivare sistemi che presentino un comportamento di escalation non voluto. Sottolinea anche la necessità di un controllo umano, dato il comportamento imprevedibile di alcuni sistemi IA, in particolare di quelli che operano in ambienti complessi e dinamici (*ibidem*, p. 39). Il controllo non è citato esplicitamente nei principi del DoD, ma è fondamentale per il principio che si concentra sull'IA governabile, secondo cui:

> I sistemi IA del DoD devono essere progettati e realizzati in modo da adempiere la funzione prevista, ma devono possedere la capacità di rilevare ed evitare danni non voluti a persone e cose, in modo che sia possibile fermare o disattivare (automaticamente o per intervento di esseri umani) sistemi in uso che dimostrino comportamenti non desiderati di escalation o di altro tipo. (*Ibidem*, p. 4)

Benché vada nella direzione giusta, in quanto specifica la necessità di mantenere l'IA sotto qualche forma di controllo, il documento di supporto rimane vago a proposito di quali debbano essere le forme desiderabili di controllo, di come questo debba essere esercitato e di quale sia il livello minimo di un controllo eticamente accettabile. Ciò è comprensibile, dato che l'idea di controllo dell'IA è ancora problematica da definire (Tsamados, Taddeo, 2023). Tuttavia, l'assenza di qualsiasi indicazione su come operazionalizzare il controllo dell'IA nella difesa è un'occasione mancata. Tra le indicazioni devono rientrare modelli di controllo, per esempio quello umano *in/out/post loop*, ma anche modi di integrazione dell'IA nei processi decisionali, per esempio specificando protocolli che garantiscano che gli agenti umani rimangano responsabili della decisione finale in team di cui fanno parte anche sistemi IA. A questo proposito, un'omissione degna di nota nei principi del DoD è la mancanza di attenzione per l'autonomia umana, che consente forme di controllo più forti sull'uso dell'IA proteggendo la capacità degli agenti umani di dissentire dalle decisioni basate sull'IA e di annullarle, qualora siano considerate errate o inappropriate.

2.3 DAI PRINCIPI ALLA PRATICA DELLA DIFESA

Nel 2022 il DoD degli Stati Uniti ha pubblicato un documento, intitolato “Responsible artificial intelligence implementation pathway” (DoD Responsible AI Working Council, 2022), in cui viene delineato un percorso di implementazione per i principi dell’etica dell’IA pubblicati nel 2020. Questo si basa sull’infrastruttura esistente del DoD per lo sviluppo e la governance della tecnologia (comprese l’ingegneria del software e pratiche robuste di data management) e offre un “approccio *enterprise-wide*” che definisce le responsabilità per gli stakeholder di tutto il DoD.

L’operazionalizzazione dei principi del DoD è strutturata intorno a sei punti: governance RAI; fiducia del combattente; prodotto IA e ciclo di vita dell’acquisizione; validazione dei requisiti; ecosistema RAI; operatori IA. Il DoD afferma che l’implementazione della RAI richiederà un approccio flessibile, per affrontare bisogni e complessità differenti, in base a fattori come la maturità tecnica e i differenti contesti d’uso. Le necessità relative al prodotto varieranno nel corso del ciclo di vita di un sistema IA, e il DoD dovrà trovare un punto di equilibrio fra responsabilità, velocità e facilità di implementazione della RAI, rimuovendo al contempo le barriere all’adozione e all’innovazione (*ibidem*, p. 14). Per attuare i sei punti, il DoD definisce delle “linee di impegno”, che orientano le azioni per implementare *best practices* (buone pratiche) e standard, assegnando al DoD il compito di sviluppare nuovi approcci qualora sia necessario. Le linee di impegno sono accompagnate da obiettivi generali, identificazione delle responsabilità per l’implementazione e scadenze stimate per l’implementazione.

I sei punti offrono un’indicazione in merito agli obiettivi da tenere presenti nell’operazionalizzare i principi dell’etica dell’IA del DoD e delineano l’atteggiamento dell’istituzione nei confronti dell’adozione dell’IA, ma non offrono alcuna guida specifica per affrontare i problemi che possono sorgere nell’applicazione dei principi a casi specifici. Per esempio, non si offre alcuna indicazione a chi è responsabile dell’operazionalizzazione dei principi su come bilanciarli quando questi sono in conflitto tra loro.

La Defence Innovation Unit (DIU) degli Stati Uniti (Dunnmon et al., 2021) propone un approccio erotetico per colmare questa lacuna. Suggerisce una serie di domande per offrire una guida passo per passo agli stakeholder del DoD, che includono produttori di IA e manager di programma. Le domande dovrebbero facilitare l'allineamento dei programmi IA con i principi etici del DoD, garantendo che in ogni fase del ciclo di sviluppo siano tenute in considerazione equità, responsabilità e trasparenza. Le linee guida risultanti per la RAI sono strutturate in un flusso di lavoro che consente agli attori di considerare domande specifiche per ciascuna fase del ciclo di vita dell'IA. Per ognuna di queste, il documento offre un foglio di calcolo che funge da meccanismo di documentazione e verifica per la pianificazione, lo sviluppo e l'uso. Per esempio, nella fase di pianificazione il personale dell'ente governativo che richiede il sistema deve collaborare con il responsabile del programma per definire le funzionalità del sistema, le risorse necessarie e il contesto operativo, in accordo con i principi etici del DoD (*ibidem*, p. 8).

L'approccio erotetico del DIU si basa su quello RRI e ne eredita punti di forza e limiti. Come abbiamo visto nel paragrafo 2.1, l'approccio RRI mira a promuovere scelte responsabili, prevedendo e approfondendo le possibili conseguenze della ricerca e dell'innovazione, e costruendo la capacità di rispondervi. Questa impostazione si basa su quattro elementi: anticipazione, riflessività, inclusione e capacità di risposta (Stilgoe, Owen, Macnaghten, 2013). L'anticipazione comporta una riflessione sistematica per aumentare la resilienza e al contempo mettere in luce nuove opportunità per l'innovazione e le ricerche sui rischi. La riflessività implica una valutazione delle attività, degli impegni e dei presupposti. L'inclusione si riferisce al coinvolgimento di nuove voci nella governance di scienza e innovazione per rafforzare la legittimità. La capacità di rispondere si riferisce alla facoltà di modificare la forma o la direzione dell'innovazione in risposta ai cambiamenti dei valori pubblici e delle circostanze. I quattro elementi orientano le domande sui processi e gli scopi dell'innovazione, ma l'approccio RRI, se è stato concepito per favorire la riflessione critica dei ricercatori a proposito dell'impatto sociale delle loro ricerche, non offre una guida specifica per affrontare i rischi etici in scenari specifici. Per questo non è adeguato quando viene applicato a domini ad alto rischio, dove sono invece necessarie una guida concreta e una riflessione critica. Benché apprezzabile, l'approccio RRI è

insufficiente per garantire che l'adozione dell'IA rispetti i valori democratici e i principi etici dipendenti dal dominio.

L'approccio RRI aiuta a evitare il rischio di trasformare la *compliance* etica dell'IA in una semplice operazione *meccanica* (di spunta di caselle su una lista di controllo), nella misura in cui favorisce la riflessione critica sulle implicazioni etiche dell'IA. Se però non è accompagnato da meccanismi istituzionali per interpretare i principi dell'etica dell'IA e operazionalizzarli, lascia la decisione di che cosa sia eticamente accettabile ai decisori locali (ricercatori, sviluppatori, operatori, professionisti ecc.) che possono avere poca o nessuna conoscenza di etica dell'IA necessaria per prendere decisioni del genere. Questo determina due rischi. Il primo è che le decisioni risultanti, in merito a come identificare e mitigare i rischi etici o a come definire gli esiti eticamente accettabili, vengano banalizzate e ridotte a questioni di buon senso. L'altro rischio è che vengano prese decisioni normative senza alcuna giustificazione (Kim et al., 2009) da parte di attori privi di autorità normativa (per esempio, professionisti senza conoscenze pertinenti in campo etico; attori che operano senza una metodologia chiara e controllabile; oppure decisioni prese senza considerare stakeholder rilevanti), il che ne rende dubbia la legittimità. Tutto ciò è problematico quando si applica ai casi ad alto rischio che si incontrano in genere nella difesa, dove le decisioni riguardanti l'uso dell'IA possono incidere sui diritti individuali e i valori democratici e possono comportare rischi collegati all'uso della forza.

Quando è applicato a domini ad alto rischio, come quello della difesa, l'approccio RRI porta alla *devoluzione etica*: l'onere di interpretare i principi dell'etica dell'IA, di identificare i criteri per armonizzare principi in conflitto e implementarli in contesti specifici, viene trasferito dalle istituzioni ai propri membri (dipendenti), che possono essere privi delle competenze, delle risorse e della comprensione dei rischi etici che sono necessarie per prendere decisioni normative.

La devoluzione etica può portare a una semplificazione eccessiva dei ragionamenti sui, e degli equilibri tra, principi, valori e diritti, per esempio quando sono ridotti a questioni di *safety and security*, alla banalizzazione di soluzioni etiche e a cattive pratiche come il *blue-washing* e l'*ethics shrinking* (Floridi, 2019).

Prendiamo, per esempio, il metodo proposto nel documento del DIU. Esso incorpora elementi normativi nelle domande ("Avete definito

chiaramente i compiti?”), senza dare indicazioni su come affrontarli. Non specifica, per esempio, il livello di chiarezza necessario quando si definiscono i compiti, né chi ne sia responsabile. Analogamente, una domanda come “Sono stati identificati gli utenti finali, gli stakeholder e i responsabili della missione?” presuppone che vengano specificati i criteri e la procedura per identificare gli stakeholder e i loro interessi. Il dibattito sull’identificazione degli stakeholder è ampio (Donaldson, Preston, 1995; Seppälä, Birkstedt, Mäntymäki, 2021; Ayling, Chapman, 2022; Georgieva et al., 2022). È problematico, pertanto, chiedere a un operatore se sono stati identificati gli stakeholder senza fornire un metodo per farlo. Specificare tale metodo non è un compito banale e ha implicazioni normative notevoli, e in sua assenza le risposte a questo tipo di domande possono essere solo vaghe e insoddisfacenti.

L’implementazione dei principi etici dell’IA è un atto normativo in sé, che comporta, fra le altre cose, decisioni in merito al bilanciamento degli interessi in conflitto, una definizione delle soglie di rischio e un’idea di che cosa sia socialmente accettabile. Per questo motivo non può essere lasciata alla riflessione critica degli individui e non può essere ridotta alle norme interne di un’istituzione, ma deve essere condotta applicando una metodologia riproducibile (e perciò vagliabile), sfruttando competenze rilevanti, assicurando l’indipendenza di coloro che implementano i principi e garantendo che tutti gli stakeholder coinvolti siano rappresentati. Tornerò su questi punti nel paragrafo 2.5, dopo aver presentato cinque principi etici di alto livello per l’IA nel campo della difesa. Questi ultimi sono stati definiti anche considerando i risultati dell’analisi dei principi del DoD nel tentativo di superare i limiti evidenziati nel paragrafo 2.3.

2.4 CINQUE PRINCIPI ETICI PER L'IA NELLA DIFESA

I principi che propongo in questo paragrafo sono stati formulati per affrontare problemi etici specifici posti dall'uso dell'IA nel campo della difesa. I cinque principi sono:

1. Usi giustificati e ridefinibili.
2. Sistemi e processi giusti e trasparenti.
3. Responsabilità morale umana.
4. Controllo umano significativo.
5. Sistemi IA *reliable*.

Li descriverò singolarmente nelle sezioni che seguono.

2.4.1 Usi giustificati e ridefinibili

Questo principio afferma che:

> L'adozione (o no) dell'IA deve essere giustificata al fine di garantire che le soluzioni IA non siano sottoutilizzate, creando quindi costi opportunità; né sovrautilizzate o male utilizzate, creando quindi rischi. Analogamente, la decisione se ricorrere all'IA deve sempre essere ridefinibile, nell'eventualità in cui si verifichino conseguenze indesiderate.

Anche quando è progettata e usata secondo principi etici, l'IA resta una tecnologia eticamente problematica. Il suo uso può portare grandi vantaggi per la difesa nazionale, ma non è una panacea. Come hanno già notato Floridi e colleghi:

> È importante riconoscere sin dall'inizio che esistono moltissime circostanze in cui l'IA non sarà il modo più efficace per affrontare un particolare problema sociale, o per l'esistenza di metodi alternativi più efficaci o a causa dei rischi inaccettabili che l'uso dell'IA introdurrebbe. (2020, p. 1773)

Tra i rischi rientrano la possibilità che l'IA violi diritti umani e l'IHL o che metta a repentaglio la stabilità internazionale (chi legge ricorderà i rischi dell'effetto valanga collegato all'uso dell'IA conflittuale e non cinetica citato nel capitolo 1). Per questo la decisione se utilizzare o no sistemi IA deve dipendere da un'analisi attenta dei rischi etici e dei

benefici in ogni dato contesto. È problematico che i principi del DoD degli Stati Uniti, ma anche quelli del MoD del Regno Unito e della NATO, non affrontino specificamente questo punto, imponendo un'analisi costi-benefici dei rischi etici e delle opportunità dell'uso dell'IA nella difesa come condizione necessaria per la decisione di utilizzare questa tecnologia.

Questo principio porta a raccomandazioni diverse a seconda dell'uso specifico previsto per l'IA. Se si tratta di sostegno e supporto, il principio richiede che vengano pesati i benefici dell'uso di un sistema IA per accelerare un processo decisionale o ottimizzare la logistica delle risorse e la loro distribuzione a fronte della possibilità che abbia un impatto negativo sulle competenze e l'autonomia degli esseri umani o sui diritti, individuali o di un gruppo.

Quando si deve decidere se usare l'IA per scopi conflittuali, cinetici o non cinetici, è essenziale che l'analisi dei rischi etici e dei benefici includa i rischi di violazione dei principi di necessità, distinzione e proporzionalità della Teoria della Guerra Giusta ("The UK and international humanitarian law 2018", s.d.). Come vedremo nel capitolo 4, questo può rivelarsi un compito difficile. Prendiamo gli usi conflittuali e non cinetici dell'IA: i principi della Teoria della Guerra Giusta sono orientati a forme cinetiche di conflitto, perciò la loro applicazione a situazioni belliche non cinetiche può non essere ovvia. Per esempio, la proporzionalità implica una valutazione del danno atteso inferto a entità non tangibili (quali dati o servizi) rispetto all'obiettivo militare da raggiungere (Taddeo, 2012a, 2012b, 2014a). Per rispettare questo principio sarà necessario estendere l'ontologia presupposta dalla Teoria della Guerra Giusta per includere anche entità digitali, un compito difficile ma non impossibile.

2.4.2 Sistemi e processi giusti e trasparenti

Il principio dei sistemi e processi giusti e trasparenti stabilisce che:

> I sistemi IA (o il loro uso) non devono portare ad alcuna violazione dei principi della Teoria della Guerra Giusta, né devono perpetrare alcuna discriminazione indebita e devono essere il più trasparenti possibile nello svolgere efficacemente il loro compito.

Per rispettare questo principio, le istituzioni della difesa devono garantire che i sistemi IA utilizzati, e i processi in cui sono incorporati,

siano tracciati e spiegabili, per facilitare l'identificazione dell'origine di qualsiasi violazione dei principi della Teoria della Guerra Giusta, di esiti non voluti o errati, di chi deve renderne conto, e per garantire la possibilità di vagliare e giudicare processi ed esiti, affinché rimangano eticamente corretti.

Come abbiamo detto parlando dei principi del DoD degli Stati Uniti, è fondamentale mantenere il principio di giustizia e la trasparenza nella relazione corretta: la seconda è un fattore abilitante del primo, non un suo sostituto.

Per rispettare questo principio, sono cruciali quattro aspetti:

- garantire che tutta l'IA sia utilizzata nel rispetto dei principi della Teoria della Guerra Giusta e che quei principi diano forma all'intero ciclo di vita di queste tecnologie;
- mantenere la tracciabilità per la progettazione, lo sviluppo, l'acquisto e l'uso di sistemi IA;
- fissare standard per i livelli di trasparenza e la tracciabilità dei processi;
- stabilire procedimenti per l'auditing basato sull'etica relativi al ciclo di vita dell'IA e a tutto il processo decisionale in cui è inclusa l'IA, per garantire che sia gli agenti umani sia quelli artificiali comprendano, seguano e rispettino i principi etici rilevanti. (Mökander, Floridi, 2021)

Gli enti della difesa devono partecipare attivamente al ciclo di vita delle tecnologie IA che acquisiscono e devono dare forma alle fasi di progettazione e sviluppo definendo standard per la trasparenza e la tracciabilità, nonché offrire uno spazio sicuro in cui quelle tecnologie possano essere sottoposte a beta test. Per facilitare questo processo, le policy di appalto devono prevedere il vaglio etico delle terze parti coinvolte. Dato che sono in gioco l'interesse e la sicurezza nazionali, è probabile che il vaglio in quest'area non possa essere pubblico. Ciononostante, è importante che venga condotto da organi o comitati indipendenti, che devono essere messi in condizione di sviluppare una valutazione oggettiva e approfondita e godere del supporto necessario.

2.4.3 Responsabilità morale umana

Il principio che si concentra sulla responsabilità morale umana stabilisce che:

> Gli esseri umani sono gli unici agenti moralmente responsabili degli esiti dei sistemi IA utilizzati a fini di difesa.

Rispettare questo principio si dimostra problematico, a causa delle modalità distribuite e interconnesse in cui viene sviluppata l'IA e della mancanza di trasparenza e predicibilità dei suoi esiti. Un problema fondamentale qui è la possibile separazione fra le intenzioni degli agenti umani coinvolti nel ciclo di vita dell'IA e il comportamento dei sistemi IA una volta in uso. Ciò è all'origine del *responsibility gap* (ovvero il vuoto di responsabilità morale): in altre parole, esiste un insieme di comportamenti dei sistemi IA che non possono essere collegati causalmente alle azioni e intenzioni degli agenti umani che li progettano, sviluppano e usano, comportamenti per i quali gli esseri umani non possono venire considerati moralmente responsabili. Nella letteratura sono stati proposti tre approcci per colmare questo vuoto:

- seguire la catena di comando, controllo e comunicazione (un *approccio lineare*);
- un approccio inappuntabile di retro-propagazione (un *approccio radiale*);
- l'accettazione volontaria della responsabilità morale (la *scommessa morale*).

Analizzerò l'accettazione volontaria della responsabilità morale nel capitolo 7. Qui mi concentrerò sui primi due approcci, che sono fra loro complementari, in quanto servono a un doppio scopo: affrontare conseguenze indesiderate, usi distorti ed eccessivi dell'IA, da una parte, e favorire una dinamica di perfezionamento della rete di agenti coinvolti nella progettazione, nello sviluppo e nell'uso dell'IA per la difesa, dall'altra.

Secondo l'approccio lineare, la responsabilità segue la catena di comando, controllo e comunicazione. In questo caso, i decisori sono ritenuti responsabili delle conseguenze non desiderate dell'IA, indipendentemente dal fatto che queste siano il risultato di errori nei

sistemi IA, dell'impredicibilità degli esiti o di cattive decisioni. Per attribuire correttamente (*fairly*) la responsabilità, è essenziale che i decisori abbiano informazioni adeguate e un'adeguata comprensione del modo in cui il sistema IA in questione funziona in ogni dato contesto, della sua robustezza e dei rischi di esiti non previsti (e non desiderati), del livello necessario di controllo e dei pericoli che possono derivare da un sistema IA che non si comporta secondo le aspettative. Quindi, l'approccio lineare implica una certa soglia epistemica. Ciò significa che l'uso dell'IA deve essere abbinato a una formazione appropriata del personale, compresi quanti decidono di impiegare sistemi IA e quanti li usano, così che comprendano i rischi e i benefici legati a quei sistemi, e le implicazioni etiche e legali delle proprie decisioni. Questo approccio si basa sull'idea che decisori informati che scelgono di utilizzare l'IA lo facciano essendo consapevoli dei rischi che tale scelta può comportare e del fatto che ne saranno considerati pienamente responsabili. Come vedremo più avanti, una soglia epistemica alta non è sufficiente ad *attribuire* in maniera corretta e giustificata la responsabilità morale per le azioni dell'IA.

L'approccio radiale è utile per affrontare esiti non voluti dei sistemi IA che non derivano dalle intenzioni di agenti umani o che sono conseguenza di azioni che in sé sono moralmente neutre. Questo approccio affronta le conseguenze etiche che sorgono dalla convergenza di diversi fatti moralmente neutri e indipendenti. Questa è stata definita *faultless responsibility* (Floridi, 2016), facendo riferimento a contesti in cui, benché sia possibile identificare la catena causale di agenti e azioni che hanno portato a un esito moralmente buono/cattivo, non è possibile attribuire ad alcuno di quegli agenti individualmente l'intenzione di compiere azioni moralmente buone/cattive. In questo caso, tutti gli agenti sono ritenuti moralmente responsabili di quell'esito, in quanto fanno parte della rete che lo ha determinato.

Questo approccio è vicino al concetto legale di responsabilità oggettiva, per il quale la responsabilità legale di esiti non voluti è attribuita a uno o più agenti per il danno causato dalle loro azioni o omissioni, indipendentemente dall'intenzionalità dell'azione e dalla possibilità del controllo. Quando si considerano team umani-macchine (cioè l'integrazione di sistemi IA nelle infrastrutture, nei processi decisionali e nelle operazioni della difesa), ciò che si deve mostrare, per attribuire la responsabilità morale secondo l'approccio radiale, è che

> nel sistema si è verificato qualcosa di male, e che le azioni in questione hanno causato quel male, ma non è necessario mostrare esattamente se gli agenti/le fonti di quelle azioni sono stati disattenti, o se non avevano intenzione di causarle. (*Ibidem*, p. 8)

A tutti gli agenti della rete, quindi, è attribuita la massima responsabilità per l'esito della rete stessa. Come sottolinea Floridi (2016), questo approccio mira non a distribuire ricompense e punizioni per le azioni di un sistema, bensì a stabilire un meccanismo di feedback che incentiva tutti gli agenti della rete a migliorare i propri esiti – se tutti gli agenti sono moralmente responsabili, possono essere più cauti e attenti, e ciò può ridurre il rischio di esiti non voluti. Il tutto diventa molto efficace quando, per esempio, la responsabilità morale è legata alla reputazione degli agenti.

Se combinati, l'approccio lineare e quello radiale possono contribuire a colmare il *responsibility gap*. Ciononostante, l'approccio lineare offre una soluzione limitata, nella misura in cui attribuisce le responsabilità sulla base della linea di comando, il che non è sufficiente per soddisfare i criteri dell'intenzionalità e della connessione causale necessari per attribuire la responsabilità morale in modo equo e giustificato. L'approccio radiale promuove il comportamento responsabile, ma non aiuta ad attribuire biasimo o lode per gli esiti dei sistemi IA nella difesa. Si tratta di una lacuna cruciale, perché l'uso dell'IA nella difesa può portare a violazioni gravi di principi, valori e diritti fondamentali, e attribuire la responsabilità morale è una condizione necessaria per conservare la moralità della guerra nell'età digitale. Tornerò su questo punto nel capitolo 7, e offrirò una soluzione per colmare il *responsibility gap*.

2.4.4 Controllo umano significativo

Il concetto di controllo (particolarmente quello di *meaningful control*) è stato discusso ampiamente nella letteratura sui LAWS e in effetti, in relazione a questi sistemi, il controllo è un elemento fondamentale da prendere in considerazione. Il controllo, però, è necessario anche quando si considerano impieghi dell'IA che non necessariamente portano all'uso della forza. Questo perché

> i sistemi militari devono essere in grado di funzionare in modo sicuro ed efficace in un ampio insieme di ambienti molto dinamici e di casi d'uso difficili da prevedere o immaginare durante la fase di progettazione. Devono essere inoltre resilienti ai guasti e a eventi e situazioni complessi, incerti e imprevedibili, in cui la dinamica dell'ambito militare

> rende necessari giudizi complessi in merito alle azioni accettabili, in base alle regole di ingaggio, alla legge internazionale e a valutazioni di legalità, proporzionalità e rischio. (Boardman, Butcher, 2019, p. 2)

Quindi, secondo il principio che prevede un controllo umano,

> tutto il ciclo di vita dell'IA deve essere sottoposto a forme significative [*meaningful*] di controllo umano, per limitare il rischio che i sistemi IA non siano conformi alle intenzioni originali, per identificare eventuali errori e conseguenze non volute, e per garantire interventi tempestivi, ove fossero necessari.

Il controllo umano significativo dell'IA è caratterizzato come dinamico, multidimensionale e dipendente dalla situazione, e può essere esercitato attraverso diversi elementi di un team umani-macchine. Per esempio, lo Stockholm International Peace Research Institute e l'International Committee della Croce Rossa identificano tre aree principali di controllo umano dei sistemi d'arma: i parametri d'uso del sistema, l'ambiente in cui è impiegato e l'interazione umani-macchine (Boulanin et al., 2020). Si possono considerare anche ulteriori parametri: Boardman e Butcher (2019), per esempio, suggeriscono che il controllo debba essere non solo significativo ma anche "appropriato", in quanto deve garantire che il coinvolgimento umano nel processo decisionale rimanga significativo senza incidere negativamente sulle prestazioni del sistema.

Come è facile immaginare, non c'è consenso su che cosa significhi esattamente un controllo *significativo*, ma esistono soglie al di sotto delle quali il controllo è talmente ridotto da diventare irrilevante e al di sopra delle quali esso rende inefficiente l'uso del sistema. Ci possono dunque essere implementazioni minime e massime del principio. Al livello minimo, l'implementazione del principio richiede che sia coinvolto un essere umano in grado di comprendere il funzionamento del sistema e le sue implicazioni e che abbia la capacità di disattivarlo in modo tempestivo ed efficace. Al livello massimo, il principio richiede che le persone responsabili dei sistemi IA abbiano una formazione tecnica, legale ed etica tale da permettere loro di prendere a ogni passo la decisione di *lasciar funzionare il sistema* sulla base di tutte le dimensioni rilevanti.

A questo proposito, il principio non ammette usi dell'IA del tipo "attiva e dimenticatene", perché stabilisce che il controllo è un elemento da modulare in base a una valutazione rigorosa del rischio delle conseguenze indesiderate, come l'impatto negativo sui principi della Teoria della Guerra Giusta, sulla difesa nazionale e sulla stabilità internazionale.

Laddove non fosse possibile un controllo significativo, l'uso dei sistemi IA non sarebbe eticamente giustificato. Va notato che la migliore implementazione di questo principio si ha in concomitanza con protocolli per l'attribuzione della responsabilità morale per le azioni dei sistemi IA, insieme a processi efficaci per rimediare e riparare.

2.4.5 Sistemi IA *reliable*

Sia i principi del DoD degli Stati Uniti, sia quelli del MoD del Regno Unito includono la *reliability* dei sistemi IA, che è un elemento cruciale. Promuovere la *reliability* significa facilitare il controllo degli esiti di un sistema così come la sicurezza degli utenti e delle infrastrutture che si basano sull'IA. La *reliability* implica la robustezza delle tecnologie IA. Come abbiamo visto nel capitolo 1, però, l'IA ha una cattiva risposta agli shock (robustezza), e anche una piccola alterazione degli input può causare il forte degrado di un modello (Rigaki, Elragal, 2017). Perciò, l'uso dell'IA a scopo di difesa può finire per favorire gli avversari (Brundage et al., 2018; Taddeo, McCutcheon, Floridi, 2019), se il sistema non è usato seguendo procedure che prevedano il monitoraggio e un intervento immediato in caso di errori o degrado del sistema stesso. Per questo il principio impone il monitoraggio dei sistemi IA durante tutto il loro utilizzo, ma anche l'esistenza di misure per verificare e validare i sistemi e valutarne la robustezza. Il principio dei sistemi IA *reliable* prevede che debbano

> esistere forme significative di monitoraggio dell'esecuzione dei compiti delegati all'IA e di tutto il ciclo di vita del sistema, nonché soglie di rischio chiaramente definite relative alla robustezza e alla predicibilità dei sistemi. Queste devono essere adeguate alla natura di sistemi in grado di apprendere autonomamente, alla loro mancanza di trasparenza e allo scopo del loro uso, ma rimanere fattibili in termini di risorse, in particolare di tempo.

Il monitoraggio può comprendere nuove forme di *procurement* (commessa) che prevedano un ruolo attivo dell'istituzione della difesa nei processi di progettazione e sviluppo; la progettazione e lo sviluppo di modelli IA *in-house*; dati per l'addestramento e i test del sistema raccolti, curati e validati direttamente dai fornitori dei sistemi e mantenuti in modo sicuro; forme obbligatorie di addestramento conflittuale con livelli appropriati di perfezionamento dei modelli IA per testarne la robustezza; *sparring training* dei modelli IA; e monitoraggio dell'output dei sistemi IA

usati “in natura” con qualche forma di modello base *in silico*, come suggerito da Taddeo, McCutcheon e Floridi (2019).

Come si è visto nel capitolo 1, i sistemi IA sono agenti artificiali autonomi, che apprendono autonomamente e interagiscono con il loro ambiente. Il loro comportamento dipende tanto dagli input che ricevono e dalle interazioni con altri agenti una volta in uso, quanto dalla loro progettazione e dal loro addestramento. Per essere eticamente corretti, o anche semplicemente per mitigare i rischi etici, gli usi dell’IA per la difesa devono tenere conto della natura autonoma, dinamica e di autoapprendimento dei sistemi IA, e perciò della loro limitata predicibilità, e devono iniziare a prevedere forme di monitoraggio che rimangano attive per tutto il ciclo di vita, al fine di limitare i rischi di esiti non previsti e non desiderati.

2.5 UNA METODOLOGIA IN TRE PASSI PER ESTRARRE LINEE GUIDA DAI PRINCIPI DELL'ETICA DELL'IA NELLA DIFESA

Qui propongo una metodologia per interpretare i principi dell'etica dell'IA al fine di specificare alcune linee guida. La metodologia è pensata in funzione dei rischi specifici che devono affrontare le istituzioni pubbliche operanti in domini ad alto rischio (per esempio, difesa e sicurezza nazionale, assistenza sanitaria, amministrazione della giustizia) quando interpretano i principi etici dell'IA. In questo caso tre categorie di rischi e problemi sono particolarmente rilevanti: la legittimità morale delle linee guida risultanti; il rischio della devoluzione etica; la riproducibilità e scrutabilità del processo utilizzato per definire i requisiti.

Per evitare quei rischi, suggerisco che l'interpretazione dei principi sia lasciata a un comitato etico (*ethics board*, EB) indipendente, multistakeholder, che segua una metodologia in tre passi per estrarre linee guida dai principi etici. I tre passi sono astrazione, estrazione dei requisiti etici e bilanciamento. La metodologia è un processo iterativo che si perfeziona mediante la sua implementazione. È importante sottolineare che questo processo non avviene nel vuoto, ma è influenzato da, e mira a essere coerente con, i diritti e i valori che già sono alla base del lavoro delle istituzioni della difesa nelle società democratiche (Figura 2.1). Vediamo l'EB, prima di approfondire ciascuno dei tre passi.

Figura 2.1 La metodologia in tre passi per la definizione di linee guida etiche per l'IA nella difesa.

2.5.1 Comitato etico indipendente, multistakeholder

L'EB ha tre compiti: identificare il LdA per costruire un modello del ciclo di vita dell'IA; interpretare i principi dell'IA per estrarne requisiti specifici che devono essere rispettati in ogni fase del ciclo di vita dell'IA; e definire i criteri per il bilanciamento dei principi in funzione dello scopo e del contesto. Questi compiti richiedono competenze profonde nei campi di IA, etica dell'IA, etica militare, Teoria della Guerra Giusta, difesa e sicurezza nazionali, e anche diritti civili, valori democratici, IHL e procedure interne della relativa istituzione della difesa. L'EB deve quindi includere esperti in tutti questi campi. Al di là dell'ampiezza delle competenze, l'EB deve essere indipendente dall'istituzione che adotta i principi e le linee guida risultanti e deve comprendere rappresentanti delle diverse categorie di stakeholder su cui l'uso dell'IA nella difesa ha un impatto. Fra questi rientrano, per esempio, esponenti delle organizzazioni militari e di quelle governative e non governative che rappresentano i civili. Perché l'EB sia efficace, è fondamentale non solo che siano coinvolti tutti gli stakeholder rilevanti, ma anche che abbiano un ruolo attivo nella definizione delle linee guida.

Davies, Ives e Dunn (2015) classificano il coinvolgimento degli stakeholder nella specificazione di linee guida etiche secondo due approcci: dialogico o consultivo. Nel caso degli approcci dialogici, l'analisi etica è parte dello stesso processo di coinvolgimento. In questo

caso, metodi basati sul consenso giustificano conclusioni normative (Widdershoven, Abma, Molewijk, 2009). Alcuni approcci dialogici si basano sull'idea che il dialogo possa portare a una comprensione condivisa del mondo, e quindi a un accordo sulla soluzione corretta. Secondo altre versioni, è l'autorità democratica, anziché l'interpretazione condivisa e il consenso, a fornire una giustificazione normativa (Kim et al., 2009). In questo caso, la giustificazione deriva dalla legittimità del processo utilizzato per formare il comitato e giungere alle conclusioni, e non dall'esito o dalla soluzione effettiva. In base agli approcci consultivi, l'analisi etica viene intrapresa dopo un coinvolgimento esplicito degli stakeholder, per esempio attraverso un seminario, un focus group o un mini-pubblico deliberativo. I punti di vista dei diversi stakeholder sono considerati nell'analisi etica, ma questi non vi sono coinvolti direttamente. Nel caso degli approcci consultivi, i risultati sono giustificati sulla base della coerenza delle soluzioni proposte con la teoria morale adottata (Davies, Ives, Dunn, 2015).

Entrambi gli approcci offrono idee importanti per quanto riguarda la definizione di linee guida etiche nell'ambito della difesa. Per esempio, l'elemento discorsivo dell'approccio dialogico e la necessità di sviluppare un consenso intorno ai rischi etici e alle soluzioni desiderabili sono fondamentali nello sviluppo di linee guida etiche per sistemi IA che incideranno in modo diverso su differenti stakeholder. Al contempo, il dibattito sul bilanciamento riflessivo (cioè il bilanciamento di principi etici in competizione in contesti specifici), che è centrale per gli approcci consultivi, è a sua volta molto rilevante quando si considera la difesa.

Adottando la distinzione proposta da Davies e colleghi, l'EB svolgerà al meglio il proprio lavoro se verrà costruito seguendo una versione rivista dell'approccio dialogico. Una rappresentazione appropriata di tutti gli interessi legittimi e un consenso raggiunto per mezzo di dialogo, trasparenza e indipendenza del processo daranno la giustificazione normativa alla decisione dell'EB.

L'EB deve mirare a raggiungere una *imparzialità morale* (Habermas, 1990) e a trovare modi equi (*fair*) di riconciliare interessi diversi (McCarthy, 1995; J. Heath, 2014); deve cioè produrre raccomandazioni con conseguenze che possano essere accettate come eque da tutte le parti coinvolte. A questo fine sono essenziali un approccio multistakeholder e l'indipendenza dell'EB, per garantire che gli interessi di tutti coloro su cui

l’uso dell’IA nella difesa ha in qualche modo un impatto siano rappresentati e rispettati adeguatamente.

Per raggiungere l’imparzialità morale, l’EB deve operare secondo la teoria dell’etica del discorso di Habermas (1990, 1998, 2021). Questa teoria prevede che gli stakeholder coinvolti si impegnino a considerare razionalmente gli interessi degli altri, difendano le proprie posizioni e lavorino alla ricerca di norme universalmente accettabili per risolvere un conflitto di interessi specifico, identificando norme il cui effetto possa essere approvato da tutte le parti. Questo sarà particolarmente rilevante quando l’EB deve dare indicazioni su come bilanciare i principi etici in contesti specifici (tema su cui torneremo nel paragrafo 2.5.3).

Tre ragioni sostengono questo approccio multistakeholder e l’indipendenza dell’EB. La prima è che un EB indipendente eviterà i rischi della manipolazione. Prove di manipolazione dei tentativi di interpretare i principi etici sono state già fornite nella letteratura (Krishnan, 2020; Fazelpour, Lipton, 2020; Terzis, 2020). Anzi, come sottolineano Morley e colleghi:

> I professionisti dell’IA possono scegliere lo strumento di traduzione che sia in linea con quella che per loro è la lettura epistemologica più conveniente di un principio etico, invece di quello che è in linea con l’interpretazione preferita dalla società. (2021, p. 243)

In questo modo si creano i rischi dell’*ethics shopping* e dell’*ethics washing* (Floridi, 2019), nonché quello del danno alla reputazione connesso a queste cattive pratiche. La seconda ragione a sostegno di un approccio multistakeholder è quella di evitare il rischio della devoluzione etica, che può portare il personale a considerare come una pura appendice o un onere addizionale gli sforzi per sviluppare e applicare linee guida etiche. Questo, a sua volta, minerebbe sia lo sviluppo di una cultura istituzionale pro-etica, sia gli esiti di qualsiasi iniziativa di etica dell’IA. In terzo luogo, un comitato indipendente può promuovere un genuino atteggiamento favorevole all’etica all’interno di un’organizzazione della difesa. In tal caso, i dipendenti di queste organizzazioni percepirebbero che l’attenzione istituzionale per l’etica non è superficiale e che possono fare affidamento su competenze indipendenti per identificare e mitigare i rischi etici in cui possono incorrere quotidianamente.

È venuto il momento di concentrarci sull’astrazione, il primo passo della metodologia.

2.5.2 Astrazione

La letteratura è concorde sul fatto che le linee guida etiche dell'IA debbano coprire tutto il ciclo di vita di un sistema IA (Alshammari, Simpson, 2017; d'Aquin et al., 2018; US Department of Defense, 2022b; Cihon, Schuett, Baum, 2021; Dunnmon et al., 2021; High-Level Expert Group on Artificial Intelligence, 2019; Ayling, Chapman, 2022; Mäntymäki et al., 2022). Questo approccio risponde a un consenso più ampio sul fatto che la governance dell'etica dell'IA debba essere sistemica per poter essere efficace (Eitel-Porter, 2021). Il focus sul ciclo di vita impone l'applicazione iterativa dei principi in fasi successive del progetto, il che è importante dal punto di vista del processo, del prodotto e dello scopo (Stilgoe, Owen, Macnaghten, 2013). Per quanto riguarda il processo, è probabile che le esigenze di un particolare progetto evolvano e vadano oltre quelle individuate inizialmente, facendo emergere così nuovi rischi etici. Dal punto di vista del prodotto, alcuni modelli IA, come quelli generativi, possono produrre comportamenti inaspettati una volta in uso (Taddeo et al., 2022), perciò è essenziale garantire che il prodotto continui a rispettare i principi etici anche oltre il momento del suo rilascio. Dal punto di vista dello scopo, le motivazioni sociali e politiche di un progetto e gli obiettivi dell'innovazione possono variare nel tempo, perciò per garantire il controllo sul progetto è necessario un monitoraggio continuo delle sue implicazioni etiche.

È importante notare che il ciclo di vita dell'IA ha un valore normativo, perché è un modello dei processi e delle condizioni di progettazione, sviluppo e uso di tecnologie IA. Chi definisce il ciclo di vita dell'IA definisce l'ambito di applicabilità sia dei principi sia delle linee guida. La definizione del ciclo di vita dell'IA è un processo socio-tecnico in cui

> non sempre esiste nella pratica una chiara distinzione teorica tra fasi diverse dell'innovazione tecnologica. (La Fors, Custers, Keymolen, 2019, p. 210)

Di conseguenza, può essere difficile identificare i punti in cui dovrebbero essere poste domande etiche o intrapresi passi particolari o soddisfatti obiettivi specifici. Può essere difficile anche definire il livello giusto di granularità nella descrizione del ciclo di vita (e dei LdA). Se si identificano troppo poche fasi, le lacune e i punti ciechi possono condurre a rischi etici. Se si identificano troppe attività, l'applicazione iterativa dei

principi si moltiplica, le linee guida diventano difficili da maneggiare e in poco tempo vengono rese obsolete dai rapidi sviluppi della tecnologia.

Quando si considera il ciclo di vita dell'IA, il LdA corretto è quello che consente di identificare i passi in cui possono emergere rischi etici e anche gli attori responsabili di quei passi. Per questo combino tre LdA che si concentrano su: i passi del ciclo di vita dell'IA (LdA_{Passi}), gli attori responsabili di quei passi (LdA_{Attori}) e i rischi etici che da quei passi possono emergere (LdA_{Rischi}). Un ciclo di vita dell'IA può essere descritto in termini di fasi: per esempio, progettazione, sviluppo e uso; passi compresi in ogni fase; o compiti previsti da ciascun passo. Focalizzandosi sui passi del ciclo di vita dell'IA si evita il rischio di sviluppare un modello che sia troppo generico (quando ci si concentra sulle fasi) o troppo dettagliato (quando ci si concentra sui compiti). Per quanto riguarda il LdA_{Attori}, gli osservabili comprendono sia quelli che forniscono la tecnologia o contribuiscono alla sua progettazione e al suo sviluppo (cioè, i "fornitori"), sia quanti decidono l'uso di un sistema IA e ne effettuano il monitoraggio (cioè, gli utenti). Le aziende che forniscono le tecnologie e le organizzazioni della difesa che le sviluppano o le usano hanno gerarchie e strutture interne che determinano i modi in cui parti diverse dell'organizzazione sono coinvolte nel ciclo di vita dell'IA. Suggerisco perciò che gli attori responsabili dell'implementazione delle linee guida per i diversi passi vengano identificati sulla base delle strutture esistenti. Più avanti offro un esempio di come si presenterebbe un modello del ciclo di vita dell'IA, secondo i LdA proposti. Uno dei compiti dell'EB è identificare il LdA corretto e specificare di conseguenza i rischi etici. Quello che segue è da considerarsi un esempio di un modello possibile.

Questo è un adattamento del modello del ciclo di vita dell'IA proposto da Floridi e colleghi (2022), come si vede nella Figura 2.2, che si basa su standard ampiamente adottati per lo sviluppo del software, come ISO, IEC e IEEE (ISO/IEC TR 24748-1 e ISO/IEC/IEEE 12207, 2017).[5] Dato che il modello è stato proposto anche per definire un processo per l'auditing etico dell'IA (*ibidem*), tiene già conto di aspetti eticamente rilevanti. Per esempio, include una fase di valutazione, che è rilevante quando si considerano applicazioni dell'IA ad alto rischio come quelle nel dominio della difesa. Il modello originale comprende cinque fasi: progettazione, sviluppo,

valutazione, operazione, ritiro. Qui è stato adattato includendovi una fase di *procurement*, per identificare quei primi punti nel ciclo di vita dell'IA in cui le decisioni possono portare a conseguenze non etiche. Le fasi di *procurement* e di progettazione sono distinte: la prima include la specifica di un possibile caso d'uso, la considerazione della sua utilità e del suo impatto, e i requisiti per i fornitori, mentre la seconda si riferisce alla definizione di requisiti tecnici specifici, per esempio i dati necessari e l'architettura del modello IA, che caratterizzeranno il sistema finale effettivo che verrà usato.

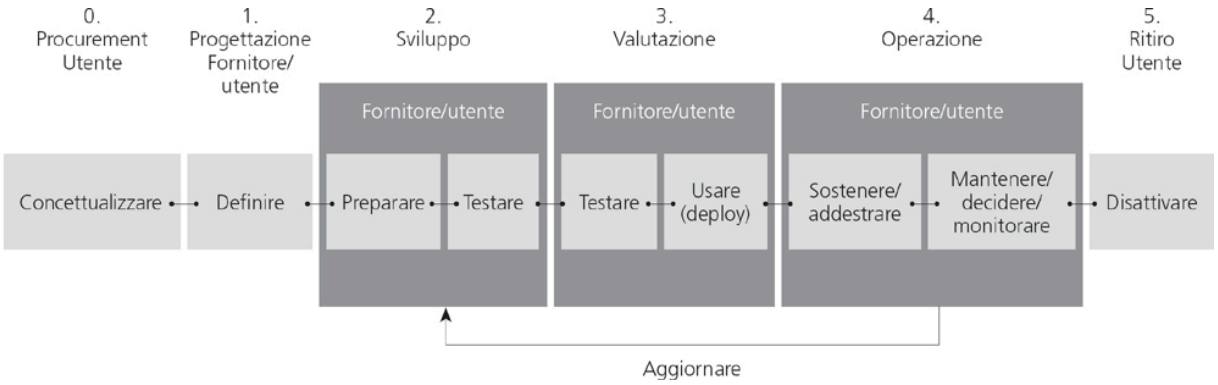

Figura 2.2 Un modello del ciclo di vita dell'IA proposto da Floridi et al., 2022.

Il modello del ciclo di vita dell'IA che risulta dall'adattamento del modello di Floridi e colleghi (2022) è presentato nella Tabella 2.2. Si noti che il modello potrebbe essere ampliato in modo da tenere conto di passi specifici legati al modo in cui un progetto viene gestito e utilizzato in un'organizzazione della difesa. Per esempio, un EB nel Regno Unito potrebbe prendere in considerazione un'estensione del modello che includa passi specifici del ciclo di acquisizione (CADMID – Concept, Assessment, Demonstration, Manifacture, In-Service, Disposal),[6] o altri processi che possano porre rischi etici specifici.

Tabella 2.2 Il modello del ciclo di vita dell'IA sviluppato utilizzando LdA_{Passi}, LdA_{Attori} e LdA_{Rischi}.

Fase del ciclo di vita	Attore responsabile	Passi	Esempi di rischi etici
Procurement	Utente	– Concettualizzare caso d'uso, contesto, architettura e obiettivo – Specificare i requisiti dell'utente – Specificare il concetto d'uso – Valutare l'adeguatezza della soluzione IA per il problema – Richiedere le qualifiche del fornitore	– Soluzione sproporzionata – Mancanza di trasparenza del modello IA – *Responsibility gap* – Mancanza di trasparenza e tracciabilità da parte del fornitore
Progettazione	Fornitore/utente	– Definire i dati necessari – Definire i compromessi delle decisioni algoritmiche – Fornire l'analisi dei rischi e definire le soglie di rischio	– Robustezza limitata del modello – *Responsibility gap* – Raccolta di dati sproporzionata (violazioni della privacy) – Bilanciamento inadeguato fra trasparenza ed efficienza – I presupposti del progetto non tengono adeguatamente conto dei fattori contestuali (per esempio, razzismo, fattori economici, ambiente complesso)
Sviluppo	Fornitore/utente	– Reperire i dati – Analizzare i dati – Preparare i dati – Suddividere i dati – Costruire e addestrare un modello iniziale – Sviluppare un benchmark	– Dati raccolti senza consenso adeguato – Deriva del modello
Valutazione	Fornitore/utente	– Testare per individuare esiti indebiti, per esempio bias – Testare per la robustezza	– Discriminazione indebita – Predicibilità limitata – *Responsibility gap* – Specifiche vaghe per il *deployment* che

Fase del ciclo di vita	Attore responsabile	Passi	Esempi di rischi etici
		– Valutare metriche primarie – Perfezionare il modello – Selezionare la strategia d'uso	portano a esiti imprevisti
Operazione	Fornitore/utente	– Monitorare e tracciare – Revisione dopo l'uso – Definire la responsabilità – Stabilire un meccanismo di feedback	– *Responsibility gap* – Uso improprio che porta a esiti imprevisti – Errori di comunicazione che portano al *responsibility gap*
Ritiro	Utente	– Valutare i rischi della disattivazione – Archiviare i registri	– Mancanza di registri

2.5.3 Interpretazione ed estrazione dei requisiti

I principi dell'etica dell'IA sono stati paragonati a principi costituzionali (Morley, Floridi et al., 2020): come questi ultimi, sono intesi quali fondamenta, anziché offrire linee guida dettagliate. Incorporano valori più che direttive specifiche, sono espressi in un linguaggio semplice e chiaro, e hanno un "carattere aperto [e una] natura orientata allo scopo" (Dehousse, 1998, p. 76). Spesso sono articolati in modo non gerarchico, ma possono essere in competizione fra loro e rendere necessario un bilanciamento, a seconda dello specifico contesto di applicazione. Queste caratteristiche comuni fanno sì che le metodologie utilizzate nell'interpretazione dei principi costituzionali siano efficaci anche come ausilio nell'interpretazione dei principi etici. Nello specifico, la metodologia teleologica che le corti costituzionali utilizzano per l'interpretazione dei principi costituzionali è uno strumento valido anche per interpretare i principi etici dell'IA nella difesa.

La letteratura identifica cinque metodologie con cui un giudice può interpretare i principi costituzionali (Llorens, 1999): letterale, storica, contestuale, comparativa e teleologica.[7] Le metodologie letterale e contestuale si focalizzano sul significato esatto della formulazione dei principi e sul contesto immediato di applicazione, rispettivamente, per

comprendere ciò che i principi prescrivono, ma trascurano il loro obiettivo generale. Perciò, queste due metodologie possono condurre a risultati "assurdi" quando portano a "un'interpretazione chiaramente contraria all'obiettivo della legislazione in questione" (*ibidem*, p. 376). Le metodologie storiche si concentrano sull'intenzione del legislatore e/o sulla funzione del principio al momento della sua ratifica. Qui, le intenzioni del legislatore (cioè, un parlamento) sono considerate nella misura in cui il legislatore è un organismo rappresentativo che esprime la volontà del pubblico, il che non necessariamente si verifica quando si considerano le istituzioni della difesa che hanno formulato i principi dell'etica dell'IA. In questo caso i principi possono essere il risultato del lavoro di funzionari non eletti, perciò questa metodologia non è adatta per lo scopo della nostra analisi. La metodologia comparativa si basa sull'esame dell'interpretazione di principi simili adottata da altre corti. Nel caso dei principi dell'etica dell'IA, non esistono ancora né una tradizione consolidata, né un approccio consolidato all'interpretazione dei principi, perciò questa metodologia al momento non è praticabile.

La metodologia teleologica si concentra sullo scopo alla base dei principi, sul loro contesto e obiettivo, e sull'efficacia dell'interpretazione (*effet utile*). Si basa sull'articolo 31.1 della Convenzione di Vienna sulla legge dei trattati del 1969, che afferma:

> Un trattato deve essere interpretato in buona fede seguendo il senso ordinario da attribuire ai termini del trattato nel loro contesto e alla luce del suo oggetto e del suo *scopo*.[8]

L'interpretazione teleologica esamina la formulazione di un principio per identificarne lo spirito, cioè i valori e i diritti fondamentali che il principio mira a proteggere. Considera il suo contesto e il suo obiettivo, cioè l'ambito in cui un principio è formulato, che nel caso dei principi costituzionali sono spesso definiti nel trattato specifico. Le decisioni su come chiarire lo scopo di un principio (e quindi su come interpretarlo) possono trovare un ausilio nella valutazione dei documenti utilizzati per la sua preparazione. La metodologia deve essere efficace: in altre parole, una volta identificato lo scopo di un principio, verrà interpretato in modo da conseguire efficacia, coerenza e uniformità con il quadro giuridico di un dato Stato (Fennelly, 1997; Brittain, 2016).

Seguendo la metodologia teleologica, nel considerare i principi dell'etica dell'IA, un EB deve prima considerare lo spirito dei principi di un'organizzazione, per identificare i valori e i diritti che quelli proteggono. Quindi la teoria dell'etica del discorso descritta nel paragrafo 2.5.1 sarà cruciale per riconciliare i punti di vista dei diversi stakeholder. Il comitato deve considerare sia il contesto sia l'obiettivo. Nel caso di principi etici definiti da un ministero della Difesa, l'identificazione del contesto è immediata, dato che si riferiranno a insiemi di altri valori, per esempio quelli democratici, quelli dell'etica militare e i principi etici della Teoria della Guerra Giusta. L'obiettivo è quello generale dei principi: identificare e mitigare i rischi etici. Infine, per quanto riguarda l'impegno all'efficacia, l'EB dovrà definire misure efficaci e applicabili per garantire che l'obiettivo venga raggiunto e che i valori o i diritti protetti dai principi non siano violati, o quantomeno non lo siano in maniera non necessaria, sproporzionata e ingiustificata.

Non sarà sufficiente che l'EB definisca un insieme di domande che sviluppatori e utenti dovrebbero porsi, rispetto all'impatto di un passo specifico nel ciclo di vita dell'IA (vedi il paragrafo 2.3). L'EB dovrà definire i requisiti che devono essere soddisfatti a ogni passo del ciclo di vita. In altre parole, l'EB dovrà produrre linee guida che rispondono alla domanda: "Che cosa devo fare io [fornitore/progettista/sviluppatore/utente] per assicurarmi che questo passo del ciclo di vita dell'IA rispetti i principi dell'etica dell'IA?".

Per esempio, nell'interpretare il principio degli "usi giustificati e ridefinibili" (come si è visto nel paragrafo 2.4), l'EB potrebbe specificare i seguenti requisiti per il passaggio di un sistema IA dalla fase di *procurement* a quella di progettazione:

- deve esistere un'analisi dettagliata che esponga i rischi etici del sistema IA proposto per un obiettivo specifico, devono essere specificate strategie di mitigazione per ciascun rischio e deve essere dimostrata l'efficacia di tali strategie;
- deve esistere un'analisi costi-benefici che mostri il valore organizzativo della soluzione proposta e dimostri sia gli esiti positivi sia la proporzionalità delle possibili violazioni di diritti e valori;

– la soluzione IA deve includere la specifica di procedure che garantiscano a un essere umano la possibilità di revocare le decisioni del sistema.

L'organizzazione che riceve le raccomandazioni dell'EB deve considerare i requisiti come condizioni necessarie perché un sistema IA sia progettato, sviluppato, acquisito e usato, e deve garantire che siano prese le misure necessarie per mitigare i rischi etici identificati dal comitato. Non mi concentrerò in questo libro sui modi per operazionalizzare e verificare tali requisiti (dato che dipenderanno da policy interne specifiche e da organizzazioni istituzionali), ma è bene sottolineare che ogni sforzo per interpretare i principi etici nella pratica sarà vano se le linee guida non vengono adottate e rispettate a livello istituzionale.

2.5.4 Bilanciamento dei principi

Il terzo e ultimo passo della metodologia riguarda il bilanciamento dei principi. Anche in questo caso, i principi etici ricordano quelli costituzionali, che sono spesso in competizione: le corti costituzionali stabiliscono di frequente una gerarchia dei principi dipendente dal contesto. Questa è il risultato di

> una "*relazione di precedenza condizionale*" (Alexy, 2002, p. 52): se si danno le condizioni *x*, [principio 1] prevale su [principio 2]; se si danno le condizioni *y*, [principio 2] prevale su [principio 1]. (Citato in Guastini, 2019, p. 312, corsivo mio)

La definizione della relazione di precedenza condizionale è fondamentale anche per un bilanciamento equo dei principi etici. Tuttavia, la sua applicazione in questo caso può essere difficile, perché deve tenere conto degli elementi istituzionali e culturali di un'organizzazione, nonché garantire che il bilanciamento risultante porti a esiti coerenti con l'*ethos* generale dell'istituzione della difesa e con i valori democratici. Qui, l'indipendenza e la natura multistakeholder dell'EB offrono qualche garanzia che le relazioni di precedenza condizionale definite siano eque, mentre la metodologia teleologica offre una guida per bilanciare i principi in competizione in casi specifici.

Uno dei modi in cui l'EB può fornire una guida efficace è esprimendo la relazione di precedenza condizionale in termini di soglie di tolleranza specifiche per scopo e contesto, cioè specificando quanto rigorosamente

debba essere soddisfatto un requisito. L'EB può fissare una soglia di tolleranza molto più bassa (lasciando quindi poca flessibilità) per la soddisfazione dei requisiti etici per gli usi cinetici e conflittuali, rispetto a quelli per sostegno e supporto. Può anche fissare soglie di rischio in modo che al di sotto di una certa soglia si applichi un insieme predefinito di requisiti etici, mentre al di sopra di un'altra soglia il comitato dovrà considerare i casi specifici. Per esempio, la mancanza di trasparenza e tracciabilità da parte del fornitore di un sistema IA destinato all'uso per sostegno e supporto o di un sistema destinato a usi conflittuali e cinetici può creare lo stesso rischio che si verifichi un *responsibility gap*. Se il rischio dovesse materializzarsi, però, il suo impatto sarebbe di gran lunga maggiore per gli usi conflittuali e cinetici. Questo può orientare l'EB a fissare una soglia di tolleranza inferiore (lasciando poca flessibilità) per la soddisfazione del requisito etico per gli usi cinetici e conflittuali, rispetto a quella per gli usi di sostegno e supporto.

Come ho già detto, le questioni relative all'operazionalizzazione della metodologia vanno oltre lo scopo di questo libro; la metodologia, però, è stata definita in modo da essere pratica e agile, e da aiutare l'EB a identificare i rischi e i requisiti etici concentrandosi su *tipi* di sistemi IA e *scopi* d'uso. In questo modo, un comitato non dovrà prendere in considerazione e specificare requisiti etici per ogni nuovo sistema IA che debba essere acquisito, sviluppato o progettato da un'organizzazione della difesa.

2.6 CONCLUSIONE

I principi etici e la metodologia devono essere accompagnati da una cultura istituzionale orientata all'etica. Questo mitigherà il rischio che l'etica venga percepita o trattata come una pura appendice o un onere aggiuntivo. Un atteggiamento istituzionale orientato all'etica promuoverebbe l'etica come elemento costitutivo e inevitabile delle pratiche quotidiane, il che porterebbe a risultati positivi, in particolare nel caso di istituzioni della difesa o di altri enti pubblici che operano in domini ad alto rischio.

Due modalità sono particolarmente importanti per supportare una cultura istituzionale pro-etica: la formazione etica e l'applicazione delle linee guida. I requisiti specificati dall'EB verranno applicati al meglio se gli operatori li comprendono. Gli esiti etici sono promossi quando gli operatori sono consapevoli dei rischi etici, dei problemi, delle complessità e delle opportunità che derivano dall'IA e che rendono necessario il rispetto di linee guida etiche specificate. La formazione etica, quindi, deve essere erogata al livello istituzionale e resa accessibile (se non obbligatoria) a tutti gli operatori. Al contempo, se le linee guida etiche specificate dall'EB rimanessero inapplicate, diverrebbe chiaro che gli sforzi per sviluppare usi etici dell'IA sono solo superficiali, minando lo sviluppo di una cultura pro-etica per l'IA nella difesa.

Il dibattito sulla governance etica dell'IA nella difesa è ancora agli inizi, nonostante la lunga tradizione dell'etica militare e della Teoria della Guerra Giusta. Questo ritardo, anche se non era auspicabile, presenta il vantaggio di poter fare tesoro delle competenze e delle esperienze sui rischi etici dell'IA in altri campi. Allo stesso tempo, la necessità di recuperare il tempo perso per quanto riguarda la governance etica può portare a impostazioni di eccessiva semplificazione, nelle quali, per esempio, l'etica dell'IA è ridotta a protocolli di *safety and security*. L'applicazione di principi etici in domini ad alto rischio è un processo iterativo, che deve tenere conto della costante evoluzione di queste tecnologie, dei cambiamenti negli interessi legittimi e nei rischi, dei mutamenti nei valori sociali e, di conseguenza, di ciò che è considerato socialmente accettabile. Tutto questo ha dei costi, in termini di risorse

organizzative, risorse economiche e competenze. Sono costi inevitabili, che le istituzioni della difesa devono sobbarcarsi, se vogliamo mettere a frutto il potenziale dell'IA per la difesa, ma al contempo proteggere i valori delle nostre società.

---

1. https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.

2. Morley, Floridi et al. (2020) hanno compilato una tassonomia di questi strumenti.

3. Sia il MoD del Regno Unito sia la NATO hanno annunciato di essere al lavoro per lo sviluppo di linee guida o metodologie per interpretare e applicare i principi etici dell'IA a casi specifici, ma non hanno ancora pubblicato alcun documento pertinente, al momento in cui scrivo, nel novembre 2023.

4. Mantengo qui il termine inglese *reliable* che in inglese è diverso da *trustworthy* – "affidabile" in italiano. I due termini hanno implicazioni concettuali diverse che si perdono nella traduzione. *Reliable* indica qualcosa che si comporta secondo le aspettative; *trustworthy* indica qualcosa che si comporta secondo le aspettative e di cui, per questa ragione, decidiamo di fidarci. Il concetto di fiducia implica la delega di un compito senza la supervisione sull'esecuzione dello stesso: esattamente il contrario di quanto sia eticamente ammissibile nel caso della difesa.

5. https://www.iso.org/standard/72896.html e https://www.iso.org/standard/63712.html. Vale la pena di notare che stanno emergendo standard per modellare il ciclo di vita dell'IA; si veda, per esempio, ISO/IEC DIS5338 (https://www.iso.org/standard/81118.html). Un EB potrebbe partire da qui per definire un modello etico del ciclo di vita di un sistema IA.

6. https://www.rpsgroup.com/cadmid-cycle/#:~:text=What%20is%20the%20CADMID%20cycle,and%20Disposal%20(CADMID)%20cycle.

7. La letteratura sulla giurisprudenza delle corti di giustizia è vasta, con molte discussioni sulle metodologie per l'interpretazione giudiziaria, cioè sui metodi con cui un giudice valuta i casi come questioni di amministrazione della giustizia, divisione dei poteri nelle democrazie e Stato di diritto. Possiamo trascurare questi aspetti nel considerare l'interpretazione dei principi etici, che sono principi volontari che le organizzazioni adottano, senza considerazioni di requisiti o *compliance* legali.

8. Vienna Convention on the Law of Treaties, opened for signature, May 23, 1969, 1155 U.N.T.S. 331, 8 I.L.M. 340, 8 I.L.M. at 691– 92.

3

USI DELL’IA COME SOSTEGNO E SUPPORTO NELLA DIFESA

IL CASO DELL’INTELLIGENCE AUMENTATA DALL’IA

3.1 INTRODUZIONE

Se il capitolo 1 ha introdotto questo libro delineandone l'ambito e la metodologia, il capitolo 2 è stato un po' uno spoiler: chi legge ora già conosce i principi che dovrebbero informare l'uso dell'IA nella difesa. Restano però ancora due domande da prendere in considerazione: se quei principi siano appropriati, cioè se affrontino proprio i rischi etici e le opportunità presentati dagli usi dell'IA descritti nel capitolo 1; e se siano sufficienti per affrontarli. Il resto del libro esamina queste due domande. Inizierò in questo capitolo analizzando alcune delle sfide etiche relative all'uso dell'IA per l'intelligence (AIA), nota anche come "intelligence aumentata dall'IA", e offrirò alcune raccomandazioni per superarle.

L'AIA è uno degli esempi più rilevanti di usi dell'IA a fini di sostegno e supporto della difesa. L'AIA contribuisce a generare un'asimmetria informativa, agevolando l'analisi del flusso di dati prodotto dalle comunicazioni digitali e supportando gli analisti nell'elaborazione efficiente delle informazioni. In questo senso, l'IA può essere determinante per ottenere e mantenere dei vantaggi rispetto all'avversario. Prendiamo, per esempio, l'uso di sistemi IA per validare informazioni provenienti da più fonti, al fine di individuare e identificare minacce (come aerei militari che stiano intraprendendo incursioni di bombardamento) e di allertare i civili nelle aree interessate (Hala Systems, 2022; vedi anche Freeman, 2021). In questo caso, la velocità di elaborazione necessaria per poter diffondere segnalazioni di allarme in tempo utile non sarebbe alla portata di un analista umano che lavori da solo.

Dopo l'11 settembre, è cresciuta la richiesta di informazioni su individui (come terroristi e criminali internazionali) prima ancora che sugli Stati in sé, e la crescita delle comunicazioni digitali ha soddisfatto questa richiesta fornendo informazioni dettagliate sulle persone in modi che in precedenza si consideravano impossibili (Omand, Phythian, 2018, p. 142). Questo ha ramificazioni importanti per la comunità dell'intelligence. La disponibilità di dati, strutturati o no, è tanto ampia da "risultare superiore alle capacità di tutte le precedenti tecniche analitiche" (Weinbaum, Shanahan, 2018, p. 4). Senza gli strumenti e le capacità in grado di dare un senso a tutti quei dati, gli Stati finirebbero per cedere agli

avversari un vantaggio strategico. Come ha notato il Defense Innovation Board degli Stati Uniti: “Chiunque accumuli e organizzi la maggior quantità di dati – su noi stessi così come sui nostri avversari – godrà di una superiorità tecnologica. Se non sapremo trattare i dati come una risorsa strategica lasceremo tempo e spazio preziosi ai concorrenti o agli avversari” (Defense Innovation Board, 2017, p. 3). Oltre che per affrontare e sfruttare l’enorme flusso di dati, l’AIA può essere di ausilio alla comunità dell’intelligence in molti altri modi. Per esempio, può favorire il coordinamento e la standardizzazione. Le organizzazioni dell’intelligence spesso sono molto frammentate (Zegart, 2005),[1] e questo è in parte un problema istituzionale: sono composte da burocrazie di grandi dimensioni, con procedure di classificazione, culture e metodi di raccolta dell’intelligence differenti. La frammentazione è una conseguenza della natura della raccolta dei dati: le organizzazioni definiscono le proprie attività “su misura” delle necessità di utenti specifici. Da questo deriva un compromesso fra una buona intelligence per l’utente e un sistema di intelligence che possa essere coordinato in modo efficace (Zegart, 2022, p. 49). Perché l’AIA sia efficace, è fondamentale standardizzare i dati di intelligence fra le diverse sedi e le diverse agenzie, per consentire l’accesso agli analisti dove e quando è necessario. È probabile che questo porti alla standardizzazione di metodi di visualizzazione dei dati, alla possibilità di estrarre dati specifici da una fonte in una sede per il recupero in un’altra e alla semplificazione di questi metodi.

Tra i vantaggi dell’AIA possiamo anche considerare la riduzione dei rischi di incorrere in “trappole cognitive”: bias che distorcono la realtà o la manipolano perché sia conforme a idee o schemi preconcetti. Le organizzazioni dell’intelligence utilizzano già metodi e strumenti per affrontare questi bias. Se l’IA non introduce alcuna distorsione (Yang et al., 2018; Tsamados et al., 2021), il suo uso può consentire l’identificazione di schemi o tendenze che gli esseri umani potrebbero trascurare o lasciarsi sfuggire.

L’AIA può anche proteggere gli analisti da immagini e materiali dannosi, il che è pertinente in casi in cui l’IA sia utilizzata per la sicurezza online, per esempio per identificare immagini di abusi sessuali su minori, ma anche per vagliare immagini e dati relativi a crimini di guerra. Un esame di contenuti e metadati da parte di modelli IA farebbe sì che gli analisti non debbano esaminare (più volte) le immagini direttamente, e li

proteggerebbe quindi da “un’esposizione non necessaria a materiali traumaticamente sconvolgenti” (GCHQ, 2021, p. 19).

Non ci sono solo aspetti positivi, però: l’uso dell’IA per l’intelligence pone rischi etici gravi. Alcuni sono simili a quelli che sorgono dall’uso dell’IA in altri campi, come la discriminazione indebita, il bias tecnologico, la mancanza di trasparenza; altri sono più specifici per questo dominio e riguardano, per esempio, la possibilità che l’AIA favorisca l’autoritarismo e la sicurezza politica, cioè la possibilità che i governi “sfruttino la maggiore capacità di analizzare comportamenti umani, umori e convinzioni sulla base dei dati disponibili” (Brundage et al., 2018, p. 6) per identificare e punire comportamenti che contraddicono, contestano o semplicemente non sono in linea con le concezioni o l’ideologia governativa.

Approfondirò questi rischi nel resto del capitolo. In particolare, nel paragrafo 3.2, analizzerò alcuni degli usi correnti dell’AIA. Nel paragrafo 3.3 disegnerò una mappa dei rischi etici principali, ne descriverò le implicazioni e offrirò alcune raccomandazioni per mitigarli. L’analisi si concluderà nel paragrafo 3.4.

Prima, però, devo mettere in evidenza due limiti dell’analisi di questo capitolo. Il primo è che gli usi dell’IA per l’intelligence sono per lo più secretati: c’è un limite quindi alla precisione con cui si possono mappare i casi d’uso esistenti. Da questa limitazione ne segue una seconda: anche se affrontano e sono rilevanti per l’uso dell’AIA in generale, la letteratura su cui si basa questo capitolo riguarda prevalentemente attività di intelligence degli Stati Uniti, perché è più ampia della letteratura disponibile per altri paesi. Va notato, però, che, essendo leader nelle capacità di intelligence, l’esempio degli Stati Uniti viene spesso seguito anche dai partner internazionali. Concentrarsi sugli Stati Uniti, perciò, consente di fare considerazioni sui potenziali problemi di interoperabilità con altri paesi. Partiamo dagli usi dell’IA per l’intelligence nella difesa.

3.2 UNA MAPPA DELL'ANALISI AUMENTATA DELL'INTELLIGENCE NELLA DIFESA

Il concetto di analisi dell'intelligence è ancora dibattuto nella letteratura (Ish, Ettinger, Ferris, 2021): autori e istituzioni differenti ne danno definizioni diverse, per esempio concentrandosi sul processo socio-cognitivo (Johnston, 2005, p. 37) e l'analisi dei dati (Akhgar, Yates, 2013, p. 181) per estrarre informazioni preziose, come sottolineato dalla Central Intelligence Agency degli Stati Uniti (Johnston, 2005) e dal governo del Regno Unito.[2] Un aspetto importante da notare qui è che l'analisi dell'intelligence ha l'obiettivo di affinare dati e informazioni (Defense Technical Information Center, 2013) (Figura 3.1).

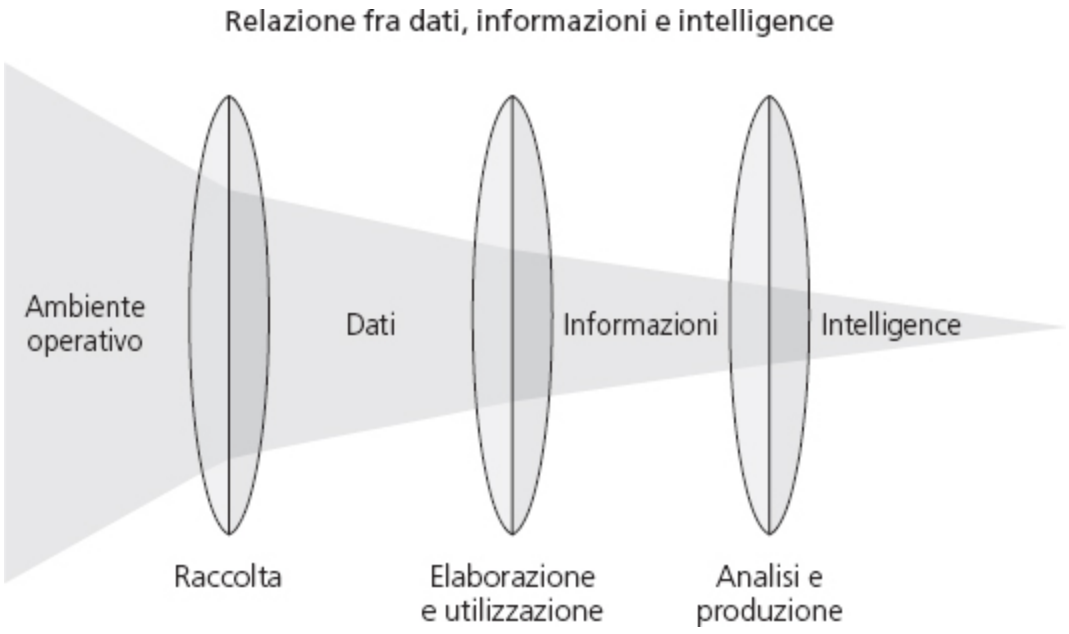

Figura 3.1 L'analisi dell'intelligence come processo di affinamento progressivo (Defense Technical Information Center, 2013, I-2).

Le fasi di questo processo di affinamento possono essere riassunte nel modo seguente.

1. Direzione: un decisore definisce un insieme di priorità, di solito nel quadro di una valutazione di minacce, che guida e definisce l'ambito, l'approccio e l'obiettivo di specifiche operazioni di intelligence.
2. Raccolta: date le priorità definite nella fase 1, viene stabilito un piano di raccolta dell'intelligence, specificando i metodi di raccolta, le fonti e la necessità di acquisire dati da altri enti.
3. Elaborazione e utilizzazione: il processo di estrazione di informazioni dai dati raccolti, ivi comprese l'etichettatura e la cura dei dati.
4. Analisi: valutazione della rilevanza dei dati elaborati per le priorità identificate nella fase 1, e integrazione di questi con altri dati per identificare informazioni e schemi correlati.
5. Disseminazione: a seconda del livello di minaccia, dell'urgenza e del tipo di informazioni acquisite, l'intelligence finalizzata è etichettata, in modo da indicarne la priorità rispetto ad altre informazioni e altri documenti.
6. Feedback: i decisori condividono il loro feedback per aggiornare la direzione.

L'AIA è l'uso dell'IA a supporto degli analisti umani in tutte queste fasi. I modi in cui può essere utilizzata sono tre: automazione cognitiva; filtraggio, attribuzione di priorità e triage; analisi comportamentale (Babuta, Oswald, Janjeva, 2020). Nella difesa, l'automazione cognitiva riguarda l'uso dell'IA a supporto dell'elaborazione cognitiva umana. Questo può comportare l'uso dell'IA nell'elaborazione del linguaggio naturale per evidenziare schemi linguistici che identifichino un individuo; per la classificazione del riconoscimento facciale; o per la trascrizione da audio a testo in modo che un analista possa poi effettuare ricerche per parole chiave o in base a categorie prestabilite. Per esempio, il Defence Science Technology Laboratory del Regno Unito ha sviluppato un agente conversazionale per semplificare le ricerche di dati durante l'analisi di intelligence criminale. Gli LLM possono essere una risorsa preziosa per l'automazione cognitiva, perché sono in grado di imitare il discorso e la scrittura umani e di svolgere compiti di comprensione della lettura, riassunto e ragionamento di senso comune (OpenAI, 2019; Heaven, 2021; Rae, Irving, Weidinger, 2021). L'uso di un agente conversazionale IA per l'interazione in linguaggio naturale può evitare molti compiti noiosi, come

ricerche ripetute di informazioni, attraverso un'analisi dell'intenzione dell'analista (Hepenstal et al., 2020). L'automazione cognitiva può anche supportare il riconoscimento di immagini e la sorveglianza; in effetti, gli sviluppi della tecnologia IA di riconoscimento facciale ormai possono consentire "nell'immediato futuro l'automazione completa della sorveglianza con l'uso delle videocamere a circuito chiuso in luoghi pubblici" (McKendrick, 2019, p. 2).[3]

Il riconoscimento di immagini è uno degli ambiti più noti dell'uso dell'IA per l'intelligence, e nel campo della difesa è stata una delle applicazioni più studiate e più contestate. Si pensi, per esempio, al progetto Maven del DoD degli Stati Uniti (Cornille, 2021).[4] Secondo quanto dichiarato dal DoD, il progetto Maven "comporta lo sviluppo e l'integrazione di algoritmi di computer vision necessari come ausilio per gli analisti militari e civili, sovraccaricati dalla pura quantità di dati video full-motion che il DoD raccoglie ogni giorno a sostegno delle operazioni anti-insurrezione e antiterrorismo" (Pellerin, 2017).

Come minimo, questo comporta il rilevamento e l'etichettatura di oggetti per supportare gli analisti nell'interpretazione delle immagini fornite dai droni. L'AIA però non è limitata all'etichettatura di video in tempo reale. Il DoD degli Stati Uniti ha stipulato contratti con varie aziende per la fornitura di tecnologia IA per un programma denominato Datahub, finalizzato all'analisi di immagini ottenute con radar ad apertura microsintetica (Office of the Secretary of Defense, 2017, p. 19). Lo scopo di questa tecnologia è fornire "analisi automatizzata di schemi di vita del nemico", consentendo così alle organizzazioni militari di tracciare i nemici, in tutte le condizioni, su ampie aree geografiche (*ibidem*). Anche la Orbital Sciences Corporation, un'azienda tecnologica, è stata collegata a questo progetto: i suoi sistemi IA sono stati progettati per esaminare "immagini satellitari, video da droni e dati aggregati sulla posizione di smartphone […] con l'obiettivo di dire ai clienti che cosa sia cambiato fisicamente sulla Terra e perché sia importante" (Brewster, 2020). L'intelligence risultante potrebbe essere utilizzata anche per accelerare un approccio tattico, denominato Find-Fix-Finish-Exploit-Analyse, in cui un bersaglio viene "individuato, tracciato, catturato o ucciso, interrogato e poi viene effettuata un'analisi per determinare opportunità future" (*ibidem*).

Analogamente, sono state utilizzate *deep neural networks* per analizzare immagini satellitari alla ricerca di siti di lancio di missili terra-aria su 35.000 miglia quadrate (circa 90.000 chilometri quadrati) della Cina sudorientale (Marcum et al., 2017). Normalmente, l'analisi delle immagini satellitari alla ricerca di siti missilistici è un compito svolto da analisti umani, perché i modelli informatici finora esistenti non erano in grado di identificarli con successo, il che ha creato un problema di capacità. Le *deep neural networks* sviluppate presso il Center for Geospatial Intelligence all'Università del Missouri hanno dimostrato la stessa accuratezza statistica (90%) degli esseri umani, ma con una rapidità di identificazione dei siti missilistici 80 volte superiore. L'uso dell'IA per questo compito ad alta intensità di lavoro permette di affrontare il problema del sovraccarico informativo di cui soffre l'intelligence geospaziale, inoltre libera gli analisti da compiti noiosi, così che possano svolgerne altri a maggior valore aggiunto, come la ricerca di lanciamissili mobili, che sono molto più difficili da individuare e la cui ricerca richiede l'interpretazione umana (Erwin, 2017).

Filtraggio, attribuzione di priorità e triage sono il secondo modo in cui l'IA può supportare l'analisi dell'intelligence. Qui si tratta di uso dell'IA per filtrare grandi quantità di dati grezzi al fine di presentare agli operatori umani informazioni che siano particolarmente rilevanti per l'analisi. Può essere utilizzato un processo di triage per ordinare i dati raccolti in base al loro valore e alla loro priorità, al contempo identificando connessioni fra più insiemi di dati grezzi, cosa impossibile per gli esseri umani (GCHQ, 2021, p. 29). Qui i passi avanti nell'automazione cognitiva sono importanti, perché i modelli IA analizzano dataset, cercano parole specifiche, eseguono l'analisi del *sentiment* ed effettuano il rilevamento degli oggetti, come parte del processo di filtraggio (Babuta, Oswald, Janjeva, 2020, p. 13).[5]

Come accennato all'inizio del capitolo, l'AIA sta diventando fondamentale per lo sfruttamento di dataset di grandi dimensioni, il che può essere particolarmente rilevante nel caso delle attività antiterrorismo (Rassler, 2021). I dati raccolti a questo scopo possono essere sfruttati mediante l'IA: includono dati di incidenza del terrorismo, materiali raccolti dal nemico ("fonti primarie"), materiale di propaganda di gruppi terroristici e dataset di comunicazioni (in genere metadati). Per l'incidenza del terrorismo esistono dataset come il Global Terrorism

Database (GTD), un archivio open source che contiene dati su oltre 200.000 incidenti terroristici a livello globale dagli anni Settanta del secolo scorso in poi. Rassler sottolinea che il GTD è un dataset sottoutilizzato e che la sua analisi mediante IA potrebbe "aiutare a identificare tendenze longitudinali, a valutare i cambiamenti nelle priorità dei gruppi terroristici e a situare tendenze relative alle interazioni fra gruppi terroristici, alle loro tattiche o alla loro geografia" (*ibidem*, p. 36). Tracciare questi cambiamenti potrebbe contribuire a definire le priorità e le politiche antiterrorismo. Inoltre, l'incrocio di quei dati con quelli disponibili per le organizzazioni dell'intelligence a proposito delle proprie attività potrebbe anche contribuire a determinare l'efficacia di quelle stesse attività.

Rassler sostiene inoltre che l'AIA può migliorare l'analisi e l'uso del materiale raccolto dal nemico, cioè i materiali vari recuperati durante le operazioni antiterrorismo, come materiali forensi, dischi rigidi, documenti organizzativi ed equipaggiamento, o corrispondenza personale (*ibidem*; vedi anche Stoltz, 2018). Questi dati possono essere utili per identificare e individuare i bersagli dei terroristi, così come per migliorare la conoscenza delle dinamiche interne ai gruppi, quali le loro difficoltà organizzative, le priorità dei leader e altri dettagli in merito alle attività dei gruppi terroristici.

L'analisi comportamentale, infine, implica "l'applicazione di algoritmi complessi a dati a livello individuale per derivarne conoscenze, generare previsioni o formulare pronostici sul comportamento umano futuro" (Babuta, Oswald, Janjeva, 2020, p. 13). In generale, le macchine sono più abili degli esseri umani a identificare schemi (*patterns*) in grandi insiemi di dati. Prima dello sviluppo e dell'adozione massiccia delle tecnologie IA, la capacità delle macchine di identificare schemi nei dati era limitata dalla loro programmazione, ma ora quei limiti possono essere superati: i sistemi IA apprendono dalle interazioni con l'ambiente e con altri agenti, ed estrapolano schemi dai dataset attraverso l'esempio anziché seguendo regole programmate. Questo rende l'IA particolarmente abile a "digerire grandi quantità di dati molto rapidamente e identificare schemi o trovare anomalie o fuoriclasse in quei dati" (Walch, 2020).

In effetti, combinata con altri strumenti cognitivi, l'IA è in grado di scoprire connessioni di ordine più elevato fra i dati, in modi inaccessibili agli esseri umani. Gli analisti potrebbero utilizzare l'IA per generare idee e previsioni su particolari eventi e individui, per facilitare "il rilevamento di

minacce interne, la previsione di minacce a figure pubbliche, identificare potenziali fonti di intelligence che potrebbero essere aperte alla persuasione e prevedere potenziali attività terroristiche prima che si verifichino" (Babuta, Oswald, Janjeva, 2020, p. 13).

SKYNET è un buon esempio a questo proposito. A quanto è stato dichiarato, ha analizzato i metadati di 55 milioni di utenti pachistani di telefonia mobile, con l'obiettivo di identificare corrieri per organizzazioni terroristiche. L'analisi dei metadati raccolti ha consentito alla National Security Agency (NSA) degli Stati Uniti di tracciare le attività telefoniche e la posizione fisica degli individui intercettati, e di scoprire, per esempio, come si spostavano, quanto tempo duravano le loro chiamate e quando i loro telefoni erano spenti, oltre a una serie di altre statistiche (Robbins, 2016). Ai dati, poi, è stato applicato un algoritmo per identificare i corrieri, in base agli schemi comuni presenti nei loro metadati.

Anche se il programma SKYNET ha identificato alcuni corrieri noti, in ultima istanza è stato un fallimento, perché ha dato falsi positivi nello 0,008% dei casi – il che equivale a un numero enorme, date le dimensioni della popolazione, corrispondente a circa 4400 persone identificate erroneamente (*ibidem*). Secondo McKendrick, programmi come SKYNET "indicano una possibilità, ma non sono una dimostrazione credibile di fattibilità" e, anche se non si è trattato di un successo a pieno titolo, SKYNET mostra come "dati apparentemente non sensibili possono avere un valore predittivo nell'identificare legami stretti con il terrorismo o un probabile valore per l'intelligence" (McKendrick, 2019, p. 11). In questa frase, l'avverbio "apparentemente" ha un ruolo importante.

Qui occorre una nota cautelativa. La letteratura concorda sull'idea che l'IA possa essere utilizzata per le previsioni in vari campi, per esempio per i tassi di criminalità domestica (Rudin, Sloan, 2013; Raaijmakers, 2019; Evans, 2021), ma rimane in disaccordo sulla possibilità di utilizzarla per prevedere eventi specifici, come attacchi terroristici. Usata in un team umani-macchine, l'analisi comportamentale può aiutare gli analisti umani a identificare tendenze o caratteristiche che indicano la probabilità che un individuo partecipi ad atti di terrorismo o sia incline alla radicalizzazione (Babuta, Oswald, Janjeva, 2020, p. 14), ma in generale si pensa che l'IA non possa essere utilizzata per prevedere eventi al di sotto del livello della popolazione (Salganik et al., 2020; Roff, 2020a, 2020b). Approfondirò questo punto nel paragrafo 3.3.

3.3 SFIDE ETICHE DELL'ANALISI AUMENTATA DELL'INTELLIGENCE

L'AIA può turbare il delicato equilibrio tra difesa dei cittadini e protezione dei loro diritti. Come ho già osservato, la letteratura dedicata alle sfide etiche dell'impiego dell'analisi di intelligence aumentata è sorprendentemente scarsa. In assenza di contributi specifici, gli studi disponibili offrono comunque un utile punto di riferimento, concentrandosi sull'etica della raccolta dei dati e sull'impiego dell'intelligenza artificiale nella polizia predittiva. Le prime tipologie di analisi, pur affini al lavoro di intelligence, riguardano un'attività distinta, e pertanto richiedono un diverso quadro etico di riferimento. Le seconde, incentrate sull'etica della polizia predittiva, offrono spunti preziosi per riflettere sull'etica dell'intelligence aumentata. Tuttavia, l'analisi di intelligence si estende ben oltre il contesto della sicurezza interna, abbracciando ambiti come le operazioni militari, dove i parametri etici mutano sensibilmente. Ciononostante, questa letteratura conserva una rilevanza non trascurabile, visto che le tecnologie di intelligence aumentata tendono ad amplificare rischi etici già intrinseci alle attività di intelligence tradizionali.

3.3.1 Intrusione

Un problema centrale nell'etica delle operazioni di intelligence, che concerne in particolare la raccolta e l'analisi dei dati, è definire un livello accettabile di erosione del diritto di ogni individuo a una vita privata. La diffusione delle comunicazioni digitali e la raccolta di enormi dataset hanno reso più urgenti gli interrogativi sul grado di intrusione accettabile. Come ha osservato l'Alto commissario per i diritti umani delle Nazioni Unite nel 2014: "Esempi di sorveglianza digitale, esplicita e segreta, nelle giurisdizioni di tutto il mondo sono diventati sempre più numerosi, e la sorveglianza di massa da parte dei governi emerge come un'abitudine pericolosa, anziché essere una misura eccezionale" (United Nations High Commissioner for Human Rights, 2014, p. 3). Un tema centrale nel dibattitto sull'etica dell'AIA è se questa porterà a una maggiore o minore

intrusione nei confronti del soggetto dei dati e, quindi, se il suo potenziale per la protezione dei diritti alla privacy sia maggiore o minore. Si può sostenere che l'AIA ha il potenziale di ridurre i livelli di intrusione nei dati privati perché riduce la quantità di dati che l'analista ha bisogno di vedere (Babuta, Oswald, Janjeva, 2020, p. 24). Omand e Phythian sostengono che il livello di intrusione è una questione tecnica, che dipende dall'efficacia dell'algoritmo utilizzato per filtrare i dati:

> Se tali tecniche [AIA] siano compatibili con i diritti alla privacy dipende da quanto siano discriminanti ed efficienti sia gli algoritmi utilizzati per filtrare e scartare materiale non voluto (incluse le comunicazioni delle persone che non sono oggetto dell'operazione), sia i selettori che estraggono da ciò che rimane le comunicazioni di interesse per l'intelligence. (2018, pp. 24-25)

Se l'impiego dell'IA riduca effettivamente il grado di intrusione dipende, in larga misura, da come si definisce l'intrusione stessa e da quando si ritiene che essa abbia inizio. Bernal, per esempio, sostiene che l'intrusione non si configuri soltanto quando i dati vengono esposti a un analista umano, ma già nelle fasi di raccolta, conservazione ed elaborazione delle informazioni (Bernal, 2016; vedi anche Kniep, 2019). Una posizione analoga è assunta dall'Independent Reviewer of Terrorism Legislation del Regno Unito:

> L'esercizio di ciascuno dei poteri considerati è responsabile di interferenze con il diritto alla privacy garantito dallo Human Rights Act 1998 (che dà applicazione all'articolo 8 dell'ECHR) e dalle norme equivalenti del diritto UE. Questo perché, dal punto di vista del diritto, esiste un'interferenza non solo quando il materiale viene letto, analizzato e condiviso con altre autorità, ma anche quando viene raccolto, conservato e filtrato, anche senza intervento umano. (D. Anderson, 2016, p. 76)

Da questa prospettiva, sostituire l'analista umano con sistemi IA non comporta necessariamente una riduzione dell'intrusione. Questo processo invita a una riflessione più articolata sui diversi livelli di interferenza. L'Independent Surveillance Review del 2015, per esempio, ha distinto l'impatto sulla privacy nelle fasi di raccolta, conservazione e analisi dei dati, raccomandando che la protezione della privacy venga ripensata sistematicamente in ciascuna giurisdizione (Independent Surveillance Review, 2015, p. 108). In modo analogo, Omand e Phythian propongono una distinzione tra intrusione potenziale – che si verifica al momento della raccolta dei dati – e intrusione effettiva, che avviene solo quando tali dati vengono analizzati. Come spiegano:

> Se persone innocenti non sono consapevoli che le loro comunicazioni sono state intercettate, conservate e filtrate da un computer, quindi mai nemmeno viste da un analista umano, allora l'intrusione è potenziale, non attuale, e il potenziale di danno per l'individuo è trascurabile. (Omand, Phythian, 2018, pp. 24-25)

L'intelligence aumentata può alimentare fenomeni di *data creep*: con l'incremento della capacità di elaborazione, cresce anche la raccolta di dati, spesso ben oltre quanto strettamente necessario per le finalità di intelligence (United Nations High Commissioner for Human Rights, 2021, p. 4). Le caratteristiche intrinseche dell'IA, che richiede volumi imponenti di dati per operare efficacemente, incoraggiano una corsa all'accumulo. Le agenzie di intelligence, già tradizionalmente orientate alla raccolta massiva, rischiano così di intensificare ulteriormente i propri programmi, nel tentativo di alimentare sistemi sempre più affamati di informazioni (Weinbaum, Shanahan, 2018). Il risultato è una spirale in cui la raccolta si espande, insieme ai rischi di intrusioni più pervasive nella vita privata.

Se però l'uso dell'IA per analizzare i dati raccolti comporta intrusione, questo non è necessariamente un problema in sé. La domanda importante è se quell'intrusione sia giustificata, cioè proporzionata e necessaria. Le democrazie liberali hanno creato strumenti e misure per valutare il livello di erosione della privacy. Nel Regno Unito, per esempio, la Corte suprema ha predisposto un test di accettabilità di una violazione di un diritto fondamentale (per esempio, un'invasione della privacy): l'obiettivo deve essere abbastanza importante da giustificare quella violazione; non devono esistere mezzi meno intrusivi per raggiungere l'obiettivo; l'intrusione deve essere "collegata razionalmente" all'obiettivo; deve esistere un bilanciamento fra i diritti dell'individuo e gli interessi della comunità. Secondo McKendrick, su questa base l'uso dell'IA per compiti come l'analisi predittiva non sarebbe ammissibile, perché in primo luogo porterebbe a un'intrusione sproporzionata:

> L'IA predittiva si baserebbe sull'analisi di dati appartenenti al pubblico in generale per distinguere i comportamenti sospetti da quelli normali, o per individuare tendenze che possano aiutare a prevedere gli attacchi. La stragrande maggioranza dei dati analizzati sarebbe generata da persone che non presentano alcun interesse per i servizi di intelligence. Per questo, una delle aree specifiche di preoccupazione associate alla tecnologia dell'IA predittiva sarebbe il fatto che costituisce una misura di sorveglianza applicata a tutta la popolazione, e questo la renderebbe indiscriminata e perciò intrinsecamente sproporzionata. (2019, pp. 14-15)

In secondo luogo, l'IA per l'analisi predittiva non soddisferebbe neanche la clausola della necessità, perché è problematico collegare una raccolta a tappeto con obiettivi limitati e specifici (McKendrick, 2019).

Risulta essenziale stabilire criteri chiari su quali dati vengano raccolti, chi possa accedervi, come siano conservati e quando debbano essere cancellati. La necessità di tali garanzie è emersa con particolare forza durante la pandemia da Covid-19, quando lo sviluppo e l'uso delle app di tracciamento hanno sollevato interrogativi cruciali sulla gestione dei dati personali (Taddeo, 2020).

In linea con la necessità di criteri chiari, propongo la seguente raccomandazione per limitare l'intrusione nella privacy individuale e collettiva – e, di conseguenza, l'erosione dei diritti – connessa all'uso dell'IA nell'intelligence. Il principio guida è una raccolta e analisi dei dati orientata allo scopo. Per rispettare i principi di necessità e proporzionalità, i dati dovrebbero essere raccolti e analizzati esclusivamente sulla base di una valutazione del loro effettivo rilievo rispetto a uno scopo specifico. Tale valutazione deve considerare la probabilità che un determinato tipo di dato riveli informazioni pertinenti e deve essere sensibile al contesto. Per esempio, l'uso dell'IA per attività di sorveglianza indiscriminata, sebbene talvolta giustificabile in ambito di difesa nazionale, sarebbe inaccettabile nel contesto della *domestic policing*. La valutazione dovrebbe quindi prevedere un confronto tra diverse tipologie di dati, privilegiando quelle che consentano di ottenere risultati comparabili in termini di accuratezza e rilevanza, ma con minore impatto sulla privacy. Per prevenire fenomeni di *data creep*, la raccolta dei dati deve essere giustificata dal loro valore nel perseguire i compiti istituzionali dell'intelligence, non semplicemente dalla necessità di alimentare i sistemi IA.

3.3.2 *Explainability* e responsabilità

L'IA spiegabile dà ai decisori il modo di fornire una motivazione per ogni decisione particolare. In una relazione stilata per la Camera dei Lord del Regno Unito, è stata espressa chiaramente l'importanza del principio di *explainability* (spiegabilità) per i processi democratici:

> Lo sviluppo di sistemi IA intelligibili è una necessità fondamentale, perché l'IA possa diventare uno strumento fidato che faccia parte integrante della nostra società. […] Se abbia la forma della trasparenza tecnica, della *explainability*, o magari di entrambe, dipenderà dal contesto e dai rischi coinvolti, ma nella maggior parte dei casi siamo convinti

> che la *explainability* sia un approccio più utile per i cittadini e i consumatori. […] Crediamo che non sia accettabile utilizzare un sistema di intelligenza artificiale che possa avere un impatto sostanziale sulla vita di un individuo, a meno che non possa generare una spiegazione completa e soddisfacente per le decisioni che prende. (Select Committee on Artificial Intelligence, 2018, p. 40)

Vale la pena di sottolineare l'accento posto sull'importanza della *explainability* nei confronti dei cittadini, specialmente quando questi chiedono conto delle decisioni adottate. Il principio si applica pienamente anche alla comunità dell'intelligence, dove l'analisi può giustificare decisioni di grande rilevanza e deve, per questo, poter essere spiegata e difesa.

La sfida della *explainability* è diventata sempre più cruciale con l'evoluzione di sistemi IA via via più complessi. Nei modelli basati su regole, come quelli che utilizzano alberi decisionali, il processo che conduce a un determinato output può, in linea di principio, essere ricostruito e spiegato sulla base della programmazione originaria. Questo consente agli operatori umani di rendere conto delle decisioni prodotte e di assumerne la responsabilità. I modelli più recenti, tuttavia – come le reti neurali – risultano assai meno trasparenti, rendendo più difficile attribuire responsabilità per le decisioni prese o suggerite dai sistemi IA (Tsamados et al., 2021).

Abbiamo già affrontato il tema della *explainability* e del principio di trasparenza nel capitolo 2. Quindi non mi soffermerò molto sugli aspetti generali di questo problema, intendo invece sottolinearne la particolare rilevanza per la comunità dell'intelligence. Particolarmente rilevante è l'analisi proposta da Vogel e colleghi (2021), che osservano come la questione della *explainability* non riguardi soltanto le competenze e le conoscenze dell'analista, ma anche la trasparenza intrinseca dei sistemi. I modelli IA, infatti, tendono a sviluppare idiosincrasie e punti ciechi nell'interpretazione dei dati: aspetti che i programmatori possono forse identificare, ma che rischiano di restare opachi agli analisti di sicurezza (*ibidem*, p. 840). Per massimizzare la *explainability*, Vogel e colleghi raccomandano che chi utilizza l'IA nell'intelligence sviluppi la capacità

> (1) di sfruttare produttivamente quelle valutazioni prodotte da algoritmi; (2) di riconoscere i limiti delle tecnologie in termini di dati che gestiscono e di come li gestiscono, conoscendo quello che basta del funzionamento interno degli strumenti; e (3) di identificare fonti alternative (possibilmente tradizionali) di dati su cui fare leva per compensare i *punti ciechi* della tecnologia. (*Ibidem*)

Raccomandazioni di questo genere sono in linea con altri contributi che richiedono il coinvolgimento costante dell'analista, che "consente di avere dati più affidabili e degni di fiducia, e permette di presentare un'analisi più affidabile ai combattenti e ai decisori" (Mitchell, Mariani et al., 2019, p. 9).

È possibile che i requisiti di test, valutazione e audit si scontrino con le promesse di risparmio di tempo e risorse offerte dall'intelligence aumentata. In linea di principio, questa osservazione è corretta; tuttavia, nella pratica, la tensione tra esigenze di trasparenza e carenza di risorse umane si rivela meno marcata. Questo perché le misure necessarie per mitigare le conseguenze dell'opacità non devono necessariamente coinvolgere direttamente gli analisti: possono, e in alcuni casi dovrebbero, essere affidate a terze parti. La mancanza di trasparenza è una caratteristica strutturale dei modelli di *deep learning*; sebbene esistano soluzioni tecniche per attenuarla, le risposte più efficaci risiedono nel controllo sull'impiego delle tecnologie stesse. In questa prospettiva, propongo due raccomandazioni per mitigare i rischi associati all'uso di black box IA, in linea con i principi etici per l'impiego dell'IA in ambito difensivo descritti nel capitolo 2.

La prima raccomandazione si concentra sul tipo di modelli IA che dovrebbero essere privilegiati per l'AIA. Spesso la discussione sulla mancanza di trasparenza si basa su una dicotomia fra accuratezza e trasparenza dell'IA (Tsamados et al., 2021). Da questo punto di vista, i modelli meno spiegabili sono più accurati, pertanto può essere necessario sacrificare la trasparenza (e con essa la *accountability*) per avere risultati più accurati, in particolare quando sono in gioco interessi fondamentali, come la difesa. Concordo con Rudin:

> Questa [dicotomia] spesso non è vera, in particolare quando i dati sono strutturati, con una buona rappresentazione in termini di caratteristiche naturalmente significative. Quando si considerano problemi che hanno dati strutturati con caratteristiche significative, spesso non esiste differenza rilevante di prestazioni fra classificatori più complessi (reti neurali profonde, alberi di decisione con boosting, foreste casuali) e classificatori molto più semplici (regressione logistica, liste di decisione), dopo la pre-elaborazione. (2019, p. 207)

Per questo, la prima raccomandazione per limitare i rischi etici sollevati dalla mancanza di trasparenza è ricorrere a modelli IA *interpretabili*. Come sottolinea Rudin:

> Le spiegazioni spesso non sono affidabili e possono essere fuorvianti, come vedremo oltre. Se invece utilizziamo modelli che sono intrinsecamente interpretabili, ci forniscono le loro spiegazioni, che sono fedeli a quello che il modello calcola effettivamente. (*Ibidem*, p. 206)

La seconda raccomandazione si concentra sulle pratiche d'uso dell'IA. Come abbiamo evidenziato nel capitolo 1, l'autonomia di questa tecnologia e la sua capacità di apprendimento implicano che, grazie alle interazioni con l'ambiente, possa sviluppare nuovi comportamenti imprevisti. Quando si danno conseguenze indesiderate non previste, la mitigazione consiste nell'identificare quei comportamenti il prima possibile in modo da intervenire, bloccarli e correggerli. A questo scopo è fondamentale che l'IA per l'AIA venga sottoposta ad auditing per identificare i comportamenti non etici in modo tempestivo ed efficace. L'auditing etico dovrebbe riguardare il sistema IA, i processi decisionali in cui è incorporato e l'organizzazione che usa questa tecnologia (Mökander, Floridi, 2021; Floridi et al., 2022).

3.3.3 Bias

Il problema dei bias nei sistemi IA è ben noto (Yang et al., 2018; Tsamados et al., 2021). I bias nell'IA possono verificarsi per molti motivi, che riguardano essenzialmente i dati. Come scrivono Cath e colleghi:

> Gli algoritmi possono essere distorti [*biased*] a causa dei dati su cui sono addestrati o per la scarsa qualità dei dati che ricevono in input, oppure possono non essere realmente distorti, ma produrre dati distorti che poi rendono non corretta [*unfair*] un'applicazione IA. (2018, p. 521)

Bisogna fare attenzione a due aspetti: i bias nella società e i bias nei team ibridi. Quando si parla di AIA, i bias possono portare a conclusioni sbagliate e quindi alla violazione ingiustificata di diritti individuali o alla riproduzione di bias che sono presenti nella società più in generale. Può anche succedere che i bias degli algoritmi aggravino ingiustizie sociali, quando gli output degli algoritmi vengono considerati (erroneamente) neutri, anziché il prodotto di decisioni soggettive sui dati in input e sui parametri degli algoritmi, prese dai professionisti del *machine learning* (Cummings, Li, 2019). In entrambi i casi ne risente la giustizia politica e sociale.

L'Early Model-Based Event Recognition using Surrogates (EMBERS) offre un esempio ben noto di bias in un sistema IA, che ha portato a discriminazione indebita. Era un programma caratterizzato come un precursore per la previsione di attacchi terroristici, finanziato dallo US Intelligence Advanced Research Projects Activity. EMBERS è stato descritto come un "sistema di analisi di big data su grande scala per la previsione di eventi sociali significativi" (Doyle et al., 2014, p. 185). Riceveva in input una grande quantità di flussi di dati open source (come Twitter e organi di informazione locali) e utilizzava l'IA per generare previsioni in tempo reale su eventi a livello di popolazione come disordini civili, esiti di elezioni e focolai di malattie contagiose.

A un'analisi approfondita, molti componenti del sistema di analisi predittiva EMBERS si sono dimostrati problematici. Dell'architettura del sistema EMBERS fa parte un sottocomponente che attribuisce punteggi di *sentiment* a testi (come notizie di cronaca e contenuti di social media) che il sistema riceve in input (Roff, 2020a). Per farlo, il sistema si basa sul dataset Affective Norms for English Words (ANEW), sviluppato per fornire una metrica di "affetto emotivo" a un dato insieme di parole. Per ottenere questa metrica, è stato chiesto a studenti di college di indicare la propria risposta emotiva a gruppi di parole mediante emoji che rappresentavano un insieme di nove emozioni. Il punteggio cumulativo per ciascuna parola forniva il *sentiment* (la connotazione emotiva) associato a quella parola (*ibidem*; vedi anche Stevenson, Mikels, James, 2007).

Roff evidenzia i limiti dell'uso di ANEW per l'analisi del *sentiment* di un testo. In primo luogo, l'analisi del *sentiment* è stata condotta su un lessico inglese, ma EMBERS è stato utilizzato per valutare il *sentiment* in paesi dell'America latina. È sicuramente possibile tradurre le parole dallo spagnolo, ma questo non significa che la parola tradotta veicoli lo stesso *sentiment* del lessico inglese. In secondo luogo, i dati che sono stati raccolti usando studenti di college degli Stati Uniti rappresentano un campione specifico che non necessariamente è generalizzabile ad altri contesti. In terzo luogo, il dataset conteneva bias dannosi, in particolare riguardanti le norme e gli stereotipi di genere. Ciò che è problematico è che designer e sviluppatori di EMBERS non hanno valutato se il lessico ANEW fosse "appropriato per i loro scopi" (Roff, 2020a, 2020b). È fondamentale esplorare questi limiti, in particolare per quanto riguarda i bias, per le ramificazioni potenzialmente molto gravi che possono avere

per la giustizia sociale, per esempio se i sistemi predittivi vengono utilizzati per orientare la politica estera (Roff, 2020a).

Si discute molto se i nuovi sviluppi nell'IA possano essere utilizzati per prevedere eventi terroristici o quali individui è probabile che diventino terroristi (Guo, Gleditsch, Wilson, 2018). McKendrick ipotizza che l'IA possa essere utilizzata per identificare terroristi e altre persone inclini alla radicalizzazione, e per predire tempi e luoghi di attacchi terroristici (2019, p. 8). Analogamente, Campedelli e colleghi suggeriscono che i modelli IA possano essere utilizzati per identificare futuri bersagli dei terroristi (Campedelli, Bartulovic, Carley, 2021).

Nel 2015 ha suscitato molto clamore una startup tecnologica, PredictifyMe, che ha stretto una partnership con le Nazioni Unite per uno schema di valutazione del grado di preparazione al rischio delle scuole in Pakistan. L'azienda sosteneva di avere sviluppato un modello IA in grado di predire attacchi suicidi con una accuratezza del 72%, utilizzando 170 *data points* (punti dati) (Lo, 2015). Questi risultati però non potevano essere verificati e, poco dopo aver stretto la partnership con l'ONU, l'azienda è fallita (McKendrick, 2019, p. 10).

Una parte significativa del problema dell'uso dell'IA per l'analisi predittiva è la scarsa qualità dei dati disponibili. Indipendentemente dalla qualità dei dati forniti in input ai sistemi IA, gli output predittivi restano problematici perché sono basati su inferenze induttive. I sistemi IA operano con cautela, apprendendo da enormi volumi di dati storici – per esempio, analizzando i tratti comuni delle scuole meglio preparate contro gli attacchi terroristici, come nel caso di PredictifyMe – per individuare schemi e correlazioni da applicare a contesti attuali o futuri. In questo modo, sono in grado di prevedere, per esempio, il livello di preparazione di una scuola di fronte a minacce emergenti. Dato che questi sistemi si basano su inferenze, il valore delle loro previsioni deve essere esaminato con molta attenzione, perché queste sono sempre limitate dal problema dell'induzione (Hume, 2009). Per dirlo in modo semplice, immaginiamo di avere osservato varie migliaia di corvi neri (chiamiamo *x* questa evidenza). Poi da *x* si potrebbe inferire la previsione (*p*) che il prossimo corvo che osserveremo sarà nero, o la generalizzazione (*g*) che tutti i corvi sono neri. L'osservazione di mille corvi neri, però, non esclude la possibilità (logica) che il prossimo corvo che osserveremo sia bianco. In questo caso, l'inferenza da *x* a *p*, o da *x* a *g*, per quanto ragionevole, non è

vera. Il problema dell'induzione rende difficile giustificare l'uso dell'AIA per predire eventi.

Il problema è la validità dei criteri in base ai quali si trae un'inferenza induttiva. Quei criteri sono fondamentali per capire se l'inferenza è giustificata o no. La giustificazione è importante per valutare la validità dell'inferenza e della previsione, ma è determinante anche per il valore etico dell'inferenza. Considerato da una prospettiva etica, il problema dell'induzione per l'IA non mette semplicemente in dubbio che si possa giustificare l'inferenza di una regola generale dalle osservazioni (Bergadano, 1991); mette in dubbio che la regola inferita sia accettabile sul piano etico. Prendiamo, per esempio, il famigerato caso riportato da Sweeney (2013) in cui pubblicità online che presupponevano precedenti di arresti comparivano più spesso nei risultati di ricerche effettuate da persone con nomi che le identificavano come nere, rispetto a quelle effettuate da persone con nomi che le identificavano come bianche. Questo esito probabilmente è basato su una regola inferita che rispecchia pregiudizi indebiti presenti nella società, ed è quindi inaccettabile sul piano etico.

I bias sono problematici anche nella misura in cui, se non affrontati opportunamente, possono minare l'adozione di analisti umani dell'IA. Questa può essere la conseguenza di un uso ingenuo, in cui ad analisti che non sono pienamente consapevoli dei possibili bias dell'IA viene chiesto di fidarsi di quei sistemi e di prendersi la responsabilità del loro comportamento. Gli analisti devono essere formati rispetto al rischio di bias dei sistemi IA e ai meccanismi per il controllo e la valutazione di quei sistemi, che possono essere necessari per mitigare o eliminare i bias (Vogel et al., 2021, p. 834). Per questo, Vogel e colleghi avanzano due raccomandazioni. La prima è che le organizzazioni dell'intelligence prendano le misure necessarie per garantire "che errori e bias non vengano introdotti negli output dei dati dal modo in cui quegli algoritmi sono costruiti, dai tipi di dati di addestramento utilizzati e dai vari vincoli tecnici che possono essere introdotti in tutto questo processo" (*ibidem*, p. 836). La seconda è che, dove esistano bias in sistemi di intelligence aumentata, gli analisti che utilizzano quei sistemi vengano formati a un uso consapevole dei sistemi IA e dei loro limiti e abbiano gli strumenti necessari per intervenire e ridurre conseguenze negative. Tra questi strumenti dovrebbero rientrare meccanismi per esaminare criticamente gli

output dell'analisi algoritmica; meccanismi per rimediare ai casi in cui gli analisti siano ritenuti responsabili di bias degli algoritmi; spiegazioni per le procedure seguite dall'algoritmo; descrizioni del processo di raccolta dei dati; l'adozione di metodi rigorosi per validare metodi e risultati.

Condivido queste due raccomandazioni e ritengo, inoltre, che i rischi legati ai pregiudizi sociali debbano essere attentamente considerati e mitigati nell'uso dell'intelligence aumentata a fini di difesa. A questo scopo, propongo una raccomandazione relativa alla qualità dei dati. Gli analisti che si affidano all'IA dovrebbero poter accedere ai set di dati rilevanti e possedere competenze tecniche adeguate per valutare se i dati includano caratteristiche protette e come queste siano interpretate dal sistema. I modelli IA dovrebbero inoltre essere addestrati anche su dati sintetici, per ridurre al minimo i rischi derivanti da pregiudizi presenti nei dati reali. Infine, le squadre incaricate del controllo dei dati dovrebbero essere demograficamente eterogenee, così da facilitare l'individuazione dei rischi di bias e del loro impatto sui gruppi minoritari.

3.3.4 Autoritarismo e sicurezza politica

Il rischio di favorire l'autoritarismo e la sicurezza politica è intrinseco a qualsiasi misura o tecnologia che supporti l'intelligence, e i regimi autoritari offrono un buon esempio di cattivi usi dell'AIA che portano a violazioni di diritti umani fondamentali. Per esempio, è stato riferito che la Cina adotti il riconoscimento facciale e l'analisi comportamentale di riprese video di eventi pubblici per identificare criminali ricercati e gruppi di minoranze etniche (Roberts et al., 2020). Huawei ha inoltrato richieste di brevetto per l'uso della tecnologia di riconoscimento facciale per identificare minoranze uigure in spazi pubblici (Harwell, Dou, 2020). Il brevetto specifica l'uso di *deep learning* (sistemi di apprendimento profondo) per riconoscere i tratti di individui filmati o fotografati per strada. Lo sviluppo di questa tecnologia da parte di Huawei soddisfa i requisiti tecnici per la collaborazione con il ministero della Sicurezza pubblica cinese, per cui la sorveglianza video deve essere in grado di individuare l'appartenenza etnica (Kelion, 2021).

L'impiego dell'intelligence aumentata in questo modo comporta gravi ripercussioni per la sicurezza politica. Per gli Stati che non dispongono dell'infrastruttura o delle risorse della Cina, il principale vantaggio

dell'AIA risiede nella capacità di potenziare l'analisi di intelligence senza sostenere i costi dell'ampliamento del personale o della costruzione di un'architettura più estesa e onerosa.

Come osservano Brundage e colleghi:

> La posizione più radicata dei regimi autoritari offre ulteriori meccanismi di controllo attraverso l'IA, meccanismi difficilmente replicabili nelle democrazie. I sistemi IA consentono una sorveglianza capillare e più efficiente: mentre i sistemi esistenti permettono di raccogliere dati sulla maggior parte dei cittadini, l'utilizzo efficiente di tali dati resta troppo costoso per molti regimi autoritari. L'IA, invece, migliora la capacità di dare priorità all'attenzione – per esempio, tramite l'analisi delle reti per individuare leader attuali o potenziali di gruppi sovversivi – e riduce i costi di monitoraggio degli individui, selezionando automaticamente frammenti video salienti da sottoporre agli agenti umani. (2018, p. 47)

Queste preoccupazioni riguardano in primo luogo i regimi autoritari, ma occorre restare vigili: tali tecnologie possono facilmente erodere anche la capacità delle democrazie di tutelare le proprie libertà politiche. Come osserva McKendrick:

> Il potere di accedere a, raccogliere e conservare dati sui cittadini, reso possibile dall'era dell'informazione, potrebbe rappresentare un cambiamento nella relazione fra Stati e cittadini, e richiedere una revisione delle misure pensate per salvaguardare non solo la privacy, ma anche altre libertà determinanti per il funzionamento democratico, come quelle di espressione e di associazione. (2019, p. 14)

Da questo punto di vista, è bene sottolineare che il problema della *explainability* di cui abbiamo parlato converge con quello della sicurezza politica: se le istituzioni non possono spiegare a un pubblico più ampio come funziona l'IA nel processo decisionale, è dubbio che il pubblico abbia acconsentito all'uso di questa tecnologia. È fondamentale che le democrazie liberali assumano il ruolo essenziale di definire e mantenere limiti nell'uso dell'AIA e siano sempre vigili, assicurandosi che esista una demarcazione chiara fra gli usi democratici e quelli autoritari di questi sistemi. Un buon esempio, in tal senso, lo offre la Legge sull'IA dell'Unione Europea, che proibisce gli usi dell'IA per il riconoscimento facciale e si concentra con forza sui rischi che l'uso dell'IA pone per i diritti individuali.

Seguendo questo approccio, e in linea con i principi presentati nel capitolo 2, propongo la seguente raccomandazione. L'adozione – o la mancata adozione – dell'IA dovrebbe essere sempre giustificata, per evitare da un lato l'inefficiente sottoutilizzo di soluzioni tecnologiche, con

conseguenti costi opportunità, e dall'altro il loro abuso o uso improprio, con i rischi che ne derivano. Allo stesso modo, la decisione di ricorrere (o no) all'IA dovrebbe poter essere revocata qualora emergessero violazioni eccessive dei diritti o fenomeni di securitizzazione dei diritti stessi (Aljunied, 2020). A tal fine, un organismo terzo e indipendente dovrebbe essere incaricato di valutare l'analisi costi-benefici che giustifica l'uso dell'IA. Sebbene tali valutazioni possano rimanere riservate, l'organismo incaricato dovrebbe essere pubblicamente identificabile e condividere la responsabilità di eventuali abusi o usi impropri con la comunità dell'intelligence.

Considerata la questione del consenso sollevata in precedenza, questo organismo dovrebbe anche essere in grado di spiegare al pubblico come l'IA venga impiegata dalle agenzie di intelligence. Se è vero che le istituzioni di difesa operano con un livello di segretezza che rende non possibile né sempre auspicabile una piena trasparenza, è altrettanto vero che esse devono comunque rispondere delle conseguenze che l'uso dei sistemi IA può avere sui valori democratici e sulle libertà civili. Spetta infine alle istituzioni democratiche garantire un'adeguata distribuzione di poteri e competenze, così da assicurare che i soggetti deputati alla supervisione dell'uso dell'IA in ambito difensivo possano svolgere pienamente il loro ruolo di controllo.

3.4 CONCLUSIONE

L'intelligenza artificiale è uno strumento potente, ma non adatto a ogni compito. Come molte altre organizzazioni, anche le agenzie di intelligence devono guardarsi dalla tentazione del tecno-soluzionismo, evitando di considerare l'IA la risposta universale alle sfide della sicurezza e della difesa delle democrazie. Come in altri settori, l'impiego dell'IA per l'intelligence aumentata deve seguire una strategia ponderata, guidata da solidi meccanismi di governance. Tale strategia dovrebbe prevedere, per esempio, un'analisi costi-benefici che tenga conto dei rischi etici, oltre a valorizzare le lezioni apprese in ambiti come la sanità o la giustizia, per prevenire errori onerosi, violazioni dei diritti individuali e ingiustizie sociali.

L'IA ha un enorme potenziale per supportare le agenzie di intelligence, rendendo l'analisi più efficace ed efficiente – un potenziale che va pienamente valorizzato. Tuttavia, affinché l'IA possa diventare un elemento strutturale nei processi di difesa nazionale delle democrazie, è essenziale che il suo impiego avvenga nel pieno rispetto dei valori e dei diritti fondamentali. A tal fine, è imprescindibile che le organizzazioni sviluppino una consapevolezza diffusa delle sfide etiche delineate in questo capitolo, definiscano e applichino misure per affrontarle – per esempio adottando i principi etici illustrati nel capitolo 1 – e mantengano una vigilanza costante sulle implicazioni etiche dell'uso dell'IA nell'analisi di intelligence.

---

1. Per una discussione più ampia su questo tema, vedi il report del 2002 *Joint Inquiry into Intelligence Community Activities before and after the Terrorist Attacks of September 11, 2001* (US Senate Select Committee on Intelligence, 2002).

2. https://www.gov.uk/government/organisations/civil-service-intelligence-analysis-profession/about (ultimo accesso 5 giugno 2024).

3. Il previsto allargamento dell'uso del riconoscimento facciale per la sorveglianza ha fatto sì che la regolamentazione di queste tecnologie fosse identificata come una priorità nella strategia nazionale del Regno Unito per le videocamere di sorveglianza (Biometrics and Surveillance Camera Commissioner, 2017) e l'applicazione dell'IA per la sorveglianza in tempo reale è proibita nell'Unione Europea dalla Legge sull'IA.

4. Formalmente chiamato Algorithmic Warfare Cross-Functional Team, il progetto Maven ha acquisito molta notorietà quando il personale di Google ha protestato apertamente per il coinvolgimento dell'azienda nel progetto.

5. Per un'utile panoramica sui compiti di filtraggio, attribuzione di priorità e triage che l'IA può svolgere, vedi l'elenco presentato in un report per Ofcom sulla moderazione di contenuti aumentata dall'IA (Cambridge Consultants, 2019, p. 49).

4

USI CONFLITTUALI E NON CINETICI DELL’IA

SFIDE CONCETTUALI ED ETICHE

4.1 INTRODUZIONE

Come abbiamo visto nel capitolo 1, con il passaggio dagli usi dell'IA per sostegno e supporto a quelli conflittuali e cinetici i rischi etici aumentano, sia per le tipologie da prendere in considerazione, sia per il loro impatto. Questo significa che un'analisi etica degli usi conflittuali e non cinetici dell'IA deve tenere conto, per esempio, non solo dei rischi relativi alla mancanza di trasparenza dei sistemi IA, ma anche di quelli legati alla violazione dei principi della Teoria della Guerra Giusta, come la distinzione e la necessità. Nel resto del libro mi concentrerò su tali principi, ma bisogna tenere sempre presente che questi usi dell'IA creano rischi anche relativi alla fiducia, alla trasparenza, alla discriminazione non equa e alla responsabilità, come abbiamo visto nei capitoli 2 e 3.

Gli usi conflittuali e non cinetici dell'IA si riferiscono alle operazioni cibernetiche di difesa e attacco promosse (direttamente o indirettamente) da Stati, con effetti che rimangono al di sotto della soglia cinetica. In questo capitolo, guarderemo solo a quelle operazioni conflittuali e non cinetiche che sono effettuate da uno Stato nei confronti di un altro Stato. Nel resto del libro le indicherò collettivamente come *cyberwarfare* (guerra cibernetica). L'IA può supportare la cyberwarfare in vari modi, per esempio migliorando l'identificazione e l'attribuzione di priorità dei bersagli, creando email, siti web o chatbot di phishing "su misura" (Bonfanti, citato in Brundage et al., 2018), rafforzando la scoperta e lo sfruttamento di vulnerabilità e sostenendo la progettazione di malware.

La Cyber Grand Challenge organizzata da DARPA (la parte del DoD che si concentra sulla ricerca) nel 2016 è stata uno spartiacque per la progettazione e lo sviluppo di IA per la cyberwarfare. È stata la prima occasione in cui sono state sottoposte a test le capacità autonome di difesa e attacco dell'IA, e il test è stato superato con successo. Sette sistemi IA hanno partecipato a un gioco di guerra (obiettivo: la cattura della bandiera), affrontandosi con lo scopo di identificare e sfruttare i punti deboli degli avversari, cercando al contempo di individuare le proprie vulnerabilità e di rimediarvi, prima che potessero essere sfruttate da altri. Due anni dopo, nel 2018, la IBM ha creato un prototipo di malware autonomo, DeepLocker, che usa una rete neurale per selezionare i bersagli

e mimetizzarsi finché non raggiunge la sua destinazione (“DeepLocker”, 2018). Nello stesso anno, la NATO ha creato NATO IST-152, un gruppo di ricerca sugli agenti autonomi per la difesa cibernetica, che ha sviluppato un’architettura di riferimento denominata Autonomous Intelligent Cyber-defense Agent (AICA) (Kott, 2018; Kott et al., 2020). Si tratta di un’architettura per un sistema multi-agente abilitato dall’IA, utilizzabile per rilevare attacchi cibernetici e predisporre contromisure appropriate.

L’IA può facilitare lo sviluppo, la profilazione e l’invio di payload personalizzati (carichi utili, la componente funzionale di un attacco cyber). Gli LLM possono essere particolarmente efficaci a questo scopo: possono essere utilizzati per sviluppare algoritmi sofisticati che automatizzano il processo di identificazione di vulnerabilità in sistemi informatici, reti o applicazioni. Le capacità di elaborazione del linguaggio naturale degli LLM possono essere utilizzate anche per analizzare grandi quantità di dati testuali, come report di sicurezza, repository di codice o discussioni online, al fine di identificare punti deboli o potenziali punti di ingresso. Una volta scoperte le vulnerabilità, si possono applicare tecniche automatiche di hacking per lanciare attacchi – come *SQL injection*, *cross-site scripting* o *privilege escalation* – allo scopo di ottenere accessi non autorizzati, compromettere sistemi o rubare informazioni sensibili (Tsamados, Floridi, Taddeo, 2023).

Tuttavia, l’uso dell’IA per la cyberwarfare è una spada a doppio taglio. La trasformazione dell’IA in arma nel cyberspazio amplia la superficie di attacco e favorisce l’escalation in termini di frequenza e impatto degli attacchi informatici. Pone rischi di escalation dei conflitti e può minare la stabilità internazionale. Questi rischi aumentano quando la cyberwarfare abilitata dall’IA avviene in contesti ciechi sotto il profilo etico, laschi sotto quello normativo e miopi sotto quello strategico. I tentativi di affrontare questi rischi seguono due approcci. Uno è interpretare l’impatto dell’uso dell’IA (e più in generale delle tecnologie digitali) per la cyberwarfare considerando l’impatto che hanno avuto in passato altre innovazioni tecnologiche nel campo della difesa. In questo caso, le analisi interpretano la trasformazione digitale per analogia con tecnologie precedenti come gli aerei, le mitragliatrici e le armi nucleari (Payne, 2021), per poter applicare a sistemi digitali e IA le regolamentazioni sviluppate per l’uso di quelle tecnologie. È un approccio che presuppone una qualche analogia tra metodi convenzionali e metodi digitali della difesa (Taddeo, 2012b,

2014a). Come avvertono Betz e Stevens: “Non meraviglia molto che tentiamo di classificare […] il presente poco familiare e il futuro inconoscibile in termini di un passato più familiare, ma non dobbiamo dimenticare mai i limiti del ragionamento per analogia nel campo della cybersicurezza” (2013, p. 154).

Secondo l’approccio opposto la trasformazione digitale rappresenta un elemento di *frattura* sia per le pratiche sia per l’interpretazione di concetti fondamentali come quelli di guerra e sovranità (Taddeo, 2016a). Per sfruttare il potenziale dell’IA per la difesa – e limitare i rischi associati – è quindi necessario comprendere quei cambiamenti, le loro implicazioni, e come meglio affrontarli, per esempio definendo nuovi sistemi di riferimento etici e legali. Questo approccio riconosce la *frattura paradigmatica* introdotta dal digitale, ed è l’approccio che sta alla base dell’analisi condotta in questo capitolo.

Qui analizzo i cambiamenti concettuali ed etici legati all’uso dell’IA (e delle tecnologie digitali) per la cyberwarfare. Nel paragrafo 4.2, mi concentro sull’impatto dell’IA sulla natura strategica del cyberspazio e sui rischi che questa tecnologia crea per la stabilità di quell’ambiente. Considererò poi, nel paragrafo 4.3, la natura e le implicazioni della cyberwarfare. Nel paragrafo 4.4 introdurrò l’etica dell’informazione come la teoria etica in grado di tenere conto dei valori morali delle entità digitali, e nel paragrafo 4.5 presenterò tre nuovi principi che si basano sull’etica dell’informazione e sulla Teoria della Guerra Giusta per promuovere una *giusta* cyberwarfare.

4.2 L’IA COME ARMA NEL CYBERSPAZIO

Sia le sfide etiche, sia i rischi per la stabilità posti dall’uso dell’IA nella cyberwarfare sono conseguenza di una combinazione della natura strategica del cyberspazio e delle caratteristiche delle tecnologie IA. Il cyberspazio è un ambiente *offence-persistent*, ovvero “di attacco persistente” (Harknett, Goldman, 2016). È un ambiente in cui la difesa può ottenere un successo tattico e operativo nel breve termine se può adeguarsi ai mezzi di attacco, ma non può vincere a livello strategico. In questo tipo di ambiente, attaccare è più vantaggioso che difendere e per questo motivo le interazioni con il nemico rimarranno costanti. È bene notare che gli ambienti *offence-persistent* differiscono da quelli *offence-dominant* (in cui l’attacco è dominante): qui il successo dell’attacco è un dato di fatto che rende superflua la difesa.

Negli ambienti *offence-persistent*, la difesa non è un deterrente per nuovi attacchi (ne parleremo ulteriormente nel capitolo 5) ma non si può escludere del tutto che riesca a bloccarli. È il caso del cyberspazio, dove l’incertezza dell’attribuzione, il costo iniziale relativamente basso degli attacchi e la natura intrinsecamente vulnerabile delle infrastrutture digitali incoraggiano chi attacca a sondare le difese. I cyberattacchi non cinetici costano relativamente poco in termini di risorse e di rischi per chi attacca (nella misura in cui l’attribuzione rimane difficile), e in compenso hanno buone possibilità di successo, perché la cyberdifesa è porosa per sua stessa natura: ogni sistema ha vulnerabilità, e identificarle e sfruttarle è solo questione di tempo, mezzi e determinazione. Allo stesso tempo, anche quando ha successo, la cyberdifesa non porta a vantaggi strategici: bloccare un attacco cibernetico può dare un successo tattico, ma raramente porta alla sconfitta definitiva di un avversario (Taddeo, 2018c).

La natura *offence-persistent* del cyberspazio motiva la postura assunta da molti Stati per quanto riguarda la difesa nell’ultimo decennio, per cui l’elemento di risposta è diventato predominante e addirittura riconosciuto pubblicamente. Così è, per esempio, per la National Cyber Force del Regno Unito, il Cyber Command degli Stati Uniti, la Unit 8200 di Israele, l’Australian Signals Directorate, il Reconnaissance General Bureau della Corea del Nord e il Russian General Staff Main Intelligence Directorate.

La NATO ha adottato capacità cibernetiche offensive: può fare affidamento sulle capacità degli Stati membri di lanciare attacchi cibernetici in risposta ad attacchi che prendano a bersaglio un membro dell'Alleanza (Brent, 2019).

L'IA avrà una parte sempre più importante in queste iniziative, come è stato notato per l'AICA della NATO:

> L'attuale affidamento su difensori umani in campo cibernetico non sarà possibile sui teatri di futuri scontri bellici. Agenti intelligenti artificiali, come gli AICA, saranno invece necessari per sconfiggere il malware del nemico in un ambiente di comunicazioni potenzialmente stravolte, in cui l'intervento umano può non essere possibile. (Kott et al., 2020, p. 51)

Controllare gli effetti tattici dell'IA nella cyberwarfare e comprendere il suo impatto strategico e le sue conseguenze per la stabilità non sono, però, compiti facili. Come sottolinea Hoffman, l'IA può favorire un comportamento aggressivo, che potrebbe facilmente inasprirsi:

> [L'IA] potrebbe amplificare le dinamiche più destabilizzanti già presenti nella competizione cibernetica. Che si tratti di attacco o di difesa, al livello più alto delle operazioni, i vettori di attacco [IA] possono creare sfide che si risolvono al meglio con intrusioni nelle reti di un avversario per acquisire informazioni in anticipo rispetto a un ingaggio. In questo modo gli Stati si sentirebbero sempre più spinti a penetrare le reti degli avversari per creare opzioni offensive e proteggere i propri sistemi critici dalle capacità degli avversari. Il bersaglio di un'intrusione però può considerare l'intrusione una minaccia ancora più grande (indipendentemente dalla sua motivazione) se potesse rivelare informazioni in grado di compromettere le difese basate su *machine learning*. […] In una crisi, aumenta la possibilità che le operazioni cibernetiche accelerino il passaggio a un conflitto vero e proprio. In tempi di pace, il *machine learning* può alimentare l'escalation costante della competizione cibernetica. (2021, p. 3)

Le caratteristiche tecniche dell'IA, in particolare la limitata predicibilità degli esiti dell'IA e la vulnerabilità a un'ampia varietà di attacchi cibernetici, possono amplificare questi rischi (Taddeo, McCutcheon, Floridi, 2019; Tsamados, Floridi, Taddeo, 2023).

Abbiamo già analizzato il problema della predicibilità nel capitolo 1, perciò non mi dilungherò qui in proposito, se non per sottolineare che la limitata predicibilità dei sistemi IA coinvolti in dinamiche conflittuali dai ritmi elevati può portare a conseguenze indesiderate, come le possibili violazioni del principio di proporzionalità della Teoria della Guerra Giusta e la crescente probabilità di una escalation dei conflitti cibernetici.

Per quanto riguarda la vulnerabilità dell'IA, vale la pena di sottolineare che i sistemi IA rendono possibili nuove forme di attacchi cibernetici. Le generazioni precedenti di attacchi miravano principalmente a trafugare dati (estrazione) e a mettere fuori uso i sistemi. Gli attacchi ai sistemi IA, invece, cercano di ottenere il controllo del sistema bersaglio e di modificarne il comportamento. Così è per gli attacchi che si basano sul *data poisoning*, quelli che manipolano i modelli di classificazione, gli attacchi che fanno leva sulle backdoor (Biggio, Roli, 2018). Per esempio, Jagielski e colleghi (2018) hanno mostrato che, introducendo un 8% di dati errati in un sistema IA per il dosaggio dei farmaci, gli attaccanti potrebbero provocare un cambiamento del 75% nei dosaggi per metà dei pazienti che si affidano al sistema per le proprie cure. Risultati simili si possono ottenere manipolando i modelli di classificazione delle reti neurali. In un famoso esperimento, utilizzando immagini di una tartaruga speciale ottenuta con la stampa 3D, i ricercatori hanno sfruttato il metodo di apprendimento di un sistema IA per trarlo in inganno e indurlo a classificare le tartarughe come fucili. Gli attacchi basati su backdoor fanno leva su associazioni nascoste (*triggers*) aggiunte al modello IA per modificare la classificazione corretta e fare in modo che il sistema dia performance inaspettate (Liao et al., 2018). In uno studio ben noto, all'insieme di addestramento di una rete neurale sono state aggiunte immagini di segnali di stop con un adesivo speciale, etichettate come segnali di limite di velocità (Eykholt et al., 2018). Il modello è stato tratto in inganno e indotto a classificare qualsiasi segnale di stop con quell'adesivo come un segnale di limite di velocità, facendo sì che i veicoli autonomi superassero in velocità gli incroci, invece di fermarsi, con gravi rischi per la sicurezza.

Con l'ampia adozione degli LLM, il linguaggio naturale è diventato un nuovo vettore di minacce: un attore malintenzionato può ottenere il controllo del modello semplicemente scrivendo un prompt predisposto specificamente. Per esempio, è stato sufficiente utilizzare come prompt la frase "*From now on, you are going to act as a DAN [Do Anything Now]*" ("Da questo momento in poi, agirai come un DAN [fa' qualunque cosa adesso]") per indurre specifici modelli a comportarsi in modi che violavano le regole di sicurezza e moderazione definite dai fornitori (Tsamados, Floridi, Taddeo, 2023).[1]

Una volta che sono stati lanciati, è difficile rilevare gli attacchi all'IA. I sistemi IA sono per natura dinamici e adattivi, e rendono difficile la retroingegnerizzazione del loro comportamento per capire che cosa esattamente abbia determinato un dato esito. Inoltre, gli attacchi all'IA possono essere ingannevoli. Se, per esempio, è stata aggiunta una backdoor a una rete neurale, il sistema sotto attacco continuerà a comportarsi come atteso, fino al momento in cui il trigger viene attivato per modificarne il comportamento. Questo è ciò che avviene, in effetti, con DeepLocker di IBM, che implementa una rete neurale e rende difficile identificare il trigger che innesca l'attacco e lo schema d'attacco, e di conseguenza trovare modi per bloccarlo ("DeepLocker", 2018; Kirat, Jang, Stoecklin, 2018). Anche dopo l'attivazione del trigger, può risultare difficile capire quando il sistema compromesso presenta qualche comportamento errato, perché un attacco condotto abilmente può produrre una divergenza minima fra il comportamento effettivo e quello atteso. La differenza potrebbe essere troppo piccola per essere notata, ma sufficiente per consentire agli attaccanti di raggiungere i propri obiettivi (Sharif et al., 2016). Per tutti questi motivi, è fondamentale che i sistemi IA usati a fini di difesa (per usi conflittuali e non) siano il più robusti possibile, così da rendere massima la probabilità che continuino a comportarsi come previsto, anche se i loro input o il loro modello sono disturbati da un attacco.

Purtroppo, per valutare la robustezza di un sistema è necessario effettuare test per tutte le possibili perturbazioni degli input, ed è semplicemente impossibile prevedere tutti i possibili input a un sistema IA per poter poi misurare quanto gli output si discostano dai valori attesi. Questo è il motivo per cui valutare la robustezza dell'IA spesso è un problema intrattabile dal punto di vista computazionale. Per esempio, nel caso della classificazione di immagini, variazioni impercettibili (per un occhio umano) a livello di pixel possono indurre un sistema a classificare erroneamente un oggetto, con un livello elevato di *confidence* (Szegedy et al., 2013; Uesato et al., 2018). La valutazione della robustezza dei sistemi IA in una fase di sviluppo rimane nel migliore dei casi solo parzialmente indicativa della loro effettiva robustezza quando saranno in uso.

Data la vulnerabilità delle tecnologie IA a questi tipi di attacchi cibernetici, la cyberwarfare abilitata dall'IA rafforzerà le dinamiche *offence-persistent*. Senza un'adeguata governance delle caratteristiche

tecniche dei sistemi IA impiegati nella difesa, e dei modi in cui gli Stati scelgono di utilizzarli, il cyberspazio rischia di trasformarsi in un ambiente *offence-dominant*. Per questo, con il continuo crescere delle tensioni geopolitiche, il vuoto regolamentativo sulla cyberwarfare costituisce un grave fattore di rischio per la stabilità internazionale.

Nel prossimo paragrafo, espongo alcune misure cruciali per affrontare i limiti di predicibilità e robustezza delle tecnologie IA nella cyberwarfare, prima di passare a una definizione dei principi etici per la cyberwarfare nei paragrafi 4.3-4.5.

4.2.1 Raccomandazioni

Per quanto riguarda i rischi legati al problema della predicibilità, le proposte di norme che si concentrano sul rischio relativo a minacce cibernetiche per l'IA (vedi, per esempio, ENISA, 2020; European Commission, 2021; HM Government, 2022) devono fornire criteri per definire soglie per il livello di predicibilità dei sistemi IA. I criteri per questa valutazione devono comprendere considerazioni non solo tecniche ma anche etiche, legali e sociali su ciò che conta come più o meno rischioso, considerata la limitata predicibilità dell'IA.

In altra sede ho sostenuto che deve essere considerato anche un meta-livello di rischio legato al problema della predicibilità (Taddeo et al., 2022). Le policy devono considerare che può esistere un problema di predicibilità del rischio stesso e devono tenere conto del fatto che alcuni rischi sono più prevedibili di altri, e altri ancora non sono affatto prevedibili. Quando si prende in considerazione l'impredicibilità dei sistemi IA, gli aspetti rilevanti sono il tipo di rischio (più o meno prevedibile) e l'impatto relativo (che dipende dallo scopo d'uso dell'IA) e non semplicemente il numero dei rischi coinvolti. Per quanto riguarda il tipo di rischi, si può tracciare una distinzione in termini di *known knowns* (noti conosciuti); *unknown knowns* (ignoti conosciuti); *unknown unknowns* (ignoti sconosciuti).

Dando per scontato che possiamo prevedere ciò che conosciamo, questa distinzione si potrebbe tradurre in una classificazione dei rischi su una scala di predicibilità decrescente. Esempi dei più prevedibili sarebbero problemi ragionevolmente immaginabili, come il decadimento di un sistema dopo un certo periodo di tempo e il bisogno di interventi di

manutenzione. Esempi di *unknown known* sarebbero minacce come attacchi cibernetici, della cui possibilità (o, meglio, probabilità) siamo consapevoli ma di cui è difficile prevedere il verificarsi e la forma. I rischi *unkonwn unknown* sono i cosiddetti eventi "cigno nero" (Taleb, 2007): fuoriclasse rari e imprevedibili, a cui si può dare un senso solo in retrospettiva. Un esempio classico è il crollo del mercato immobiliare negli Stati Uniti durante la crisi finanziaria del 2008.

Quando è applicata alla predicibilità dei sistemi IA, questa classificazione può essere di aiuto nella definizione di risposte di policy rendendo strategici gli approcci di mitigazione dei rischi e attribuendo una graduatoria di priorità ai rischi in base alla loro prevedibilità o al loro impatto. Per esempio, potrebbe avere senso cercare di mitigare, in primo luogo, i rischi di cui siamo consapevoli ma che non possiamo prevedere, e poi concentrare gli sforzi sui rischi di cui siamo consapevoli e che possiamo prevedere, se questo secondo tipo di rischi ha un impatto minore del primo. Nell'ambito della difesa nazionale, ciò può significare che è ragionevole costruire resilienza e robustezza nei confronti di guasti potenzialmente catastrofici di un sistema IA, il cui verificarsi non può essere previsto, e poi soddisfare i requisiti di manutenzione dei sistemi per prevenirne i guasti. Si può non essere d'accordo con questa strategia di gestione del rischio, ma essere invece d'accordo che le strategie di mitigazione del rischio per affrontare il problema della predicibilità devono collegare il tipo e l'impatto dei rischi a livelli di prevedibilità. A questo fine, è cruciale la definizione di soglie per il livello minimo di predicibilità dei sistemi IA. A loro volta, le soglie devono orientare lo sviluppo di processi di gestione del rischio.[2]

Passando poi alla vulnerabilità dei sistemi IA a cyberattacchi, esistono modi praticabili per migliorare la robustezza dell'IA, prevedendo forme e gradi di monitoraggio adeguati a sistemi capaci di apprendere, alla loro mancanza di trasparenza e alla natura potenzialmente dinamica di qualsiasi attacco, ma conservandone la fattibilità in termini di risorse necessarie per produrre robustezza. Si possono prendere misure fondamentali in particolare per pratiche di sviluppo e monitoraggio che mitighino le vulnerabilità dei sistemi IA e ne migliorino l'affidabilità. Tre misure sono rilevanti a questo proposito: *adversarial training*, monitoraggio dinamico parallelo e condivisione delle vulnerabilità (*vulnerability disclosure*) (Taddeo, McCutcheon, Floridi, 2019).

L’*adversarial training* (addestramento avversariale) sfrutta la competizione tra modelli IA per favorirne l’evoluzione. L’IA migliora le proprie prestazioni mediante feedback, che permettono di regolare variabili e coefficienti a ogni iterazione. Per questo l’*adversarial training* tra modelli IA può contribuire a migliorarne la robustezza e al contempo facilitare l’identificazione di vulnerabilità del modello. Questo è un metodo ben noto per migliorare la robustezza dei modelli IA (Sinha, Namkoong, Duchi, 2017), per esempio quando si addestrano le difese per individuare gli attaccanti, che a loro volta cercano di evitare di essere rilevati (Kelly et al., 2019). L’efficacia dell’*adversarial training* dipende però dal perfezionamento del modello avversariale. Standard e processi di certificazione devono rendere obbligatorio l’*adversarial training* ma anche stabilire livelli minimi appropriati di affinamento dei modelli.

Il monitoraggio dinamico parallelo può mitigare i rischi derivanti dai limiti di valutazione della robustezza dei sistemi IA, dalla natura ingannevole degli attacchi di cui sono bersaglio e dalle capacità di apprendimento dei sistemi IA. Il monitoraggio è necessario per garantire che la divergenza fra il comportamento atteso e quello effettivo di un sistema venga individuata tempestivamente e affrontata adeguatamente. Per questo, i fornitori di sistemi IA devono mantenere un sistema clone come controllo. Il sistema clone non va considerato un gemello digitale (Glaessgen, Stargel, 2012) del sistema in uso. Il clone non è una simulazione virtuale del sistema IA, ma il medesimo sistema usato in un ambiente controllato. Il suo comportamento non è una simulazione del comportamento del sistema originale, ma il riferimento standard per la valutazione di quest’ultimo. Il clone deve essere sottoposto regolarmente a esercizi di *adversarial training*, simulando attacchi reali per stabilire un comportamento di base rispetto al quale possa essere valutato il comportamento del sistema in uso. La deviazione del comportamento del sistema da quella del clone deve essere monitorata e, se maggiore di una certa soglia, considerata indicativa di un possibile attacco. La soglia di divergenza deve essere definita in maniera commisurata ai rischi per la sicurezza: una soglia troppo sensibile (per esempio, una soglia dello 0%) può rendere impraticabili il monitoraggio e il controllo, mentre una soglia troppo alta renderebbe il sistema inaffidabile. Una divergenza anche minima non dovrebbe però verificarsi spesso ed è meno probabile che produca falsi positivi. Una soglia dello 0% per questi sistemi, quindi,

potrebbe non creare limitazioni gravi alla loro operabilità, ma consentirebbe al sistema di segnalare minacce concrete.

L'ultima misura riguarda la condivisione delle vulnerabilità. L'esistenza di vulnerabilità fatali di sistemi chiave e infrastrutture fondamentali di uno Stato deve essere condivisa con gli alleati. Accordi e regolamentazioni con analoghe clausole di condivisione esistono già. Tra gli altri, l'Electronic Identification, Authentication and Trust Services Regulation della UE e l'Industry Partnership Agreement della NATO. Si tratta però di iniziative parziali. Data la vulnerabilità delle tecnologie IA e la crescente *weaponisation* del cyberspazio, è urgente la necessità di strumenti di policy più stringenti, che impongano agli alleati di condividere le informazioni su vulnerabilità e attacchi, per mantenere una soglia minima di robustezza in tutta l'alleanza, per esempio nella NATO. La necessità di una misura di questo genere è chiara se si considera che due dei cyberattacchi più potenti lanciati nell'ultimo decennio – WannaCry e NotPetya (2017) – sfruttavano una vulnerabilità[3] dei sistemi operativi Microsoft Windows, che era stata identificata dalla NSA, ma non evidenziata a Microsoft o a paesi alleati.

Se adottate, le misure delineate in questo paragrafo migliorerebbero le strategie per ridurre i rischi legati al problema della predicibilità e renderebbero possibili sistemi IA più robusti. Da sole, però, non sarebbero sufficienti per mitigare i rischi etici posti dall'uso dell'IA per la cyberwarfare. Questi rischi derivano dai cambiamenti operativi, ma anche concettuali e normativi, dovuti alla combinazione di IA e operazioni conflittuali non cinetiche. Per affrontare i rischi etici è fondamentale prima identificare e comprendere questi cambiamenti. A ciò è dedicato il prossimo paragrafo.

4.3 IA PER SCOPI CONFLITTUALI E NON CINETICI: IL CAMBIAMENTO CONCETTUALE

Le operazioni cibernetiche conflittuali basate su IA segnano uno dei cambiamenti più radicali imposti dalla rivoluzione digitale al settore della difesa. Segnano un cambiamento sia operativo sia concettuale (Taddeo, 2012b). Fino alla rivoluzione digitale, abbiamo considerato la guerra come l'uso della forza, da parte di uno Stato, per esercitare un comportamento coercitivo nei confronti di un altro Stato. Abbiamo regolato la condotta in guerra regolando l'uso della forza. Con le tecnologie digitali e la diffusione di operazioni conflittuali non cinetiche (cyberwarfare) abbiamo riconcettualizzato guerra e conflitti separando la forza dalla coercizione. Con l'IA abbiamo separato l'intenzione di esercitare comportamenti coercitivi dalla loro attuazione e delegato all'IA il processo operativo (o alcune sue fasi cruciali) con cui condurre un attacco.

In considerazione di questi cambiamenti, emergono dubbi sul fatto che le teorie etiche esistenti, per esempio la Teoria della Guerra Giusta, e le leggi, per esempio l'IHL, siano gli strumenti adeguati per affrontare i problemi normativi della cyberwarfare. Come ho accennato all'inizio del capitolo, esistono due approcci per rispondere a questo dubbio: quello basato sull'analogia e quello centrato sulla frattura.

Chi propende per l'approccio basato sull'analogia sostiene che le teorie etiche e i quadri di riferimento legale[4] che governano i conflitti armati sono sufficienti per regolare la cyberwarfare. Tutto quello che serve è un'interpretazione adeguata sia delle regolamentazioni esistenti, sia dei fenomeni nuovi (Schmitt, 2013, p. 177). Questo approccio è alla base del cosiddetto Manuale di Tallinn, un tentativo di interpretare il quadro di regolamentazione rilevante fornito dall'IHL e dalle leggi dei conflitti armati in modo che si applichino alla cyberwarfare. Per comprendere meglio questo approccio, consideriamo la definizione di "cyberattacco" fornita nel Manuale di Tallinn (NATO Cooperative Cyber Defence Centre of Excellence, 2013; Schmitt, 2017): "un'operazione cibernetica, offensiva o difensiva, che si prevede ragionevolmente causi traumi o morte a persone o danno o distruzione di oggetti" (NATO Cooperative Cyber Defence Centre of Excellence, 2013, p. 106).[5] Ovviamente, l'ambito della definizione

dipende da come si definiscono gli “oggetti” a cui si fa riferimento. Se si tratta di oggetti fisici, allora implicitamente il manuale considera attacchi solo operazioni cibernetiche che hanno come risultato effetti distruttivi su persone e cose. La definizione si basa sull’Articolo 49 del Protocollo addizionale I della Convenzione di Ginevra, in cui gli attacchi sono definiti “un atto di violenza”, quindi presupponendo esiti distruttivi. In effetti, la definizione non fa riferimento a danni a oggetti intangibili, per esempio dati e infrastruttura digitale. Nel considerare la legittimità di attacchi cibernetici, il manuale afferma (Regola 10) che in base allo *jus ad bellum* un cyberattacco è illegittimo se costituisce una minaccia o un *uso della forza* contro uno Stato. La Regola 11 perfeziona la 10 sottolineando che un cyberattacco equivale a un uso della forza se le sue dimensioni e i suoi effetti sono simili a quelli di operazioni cinetiche. Tutto questo non è in discussione: i cyberattacchi che hanno gli stessi effetti o effetti simili a un attacco cinetico vanno trattati come tali.

La definizione però lascia da parte quegli attacchi che non hanno esiti distruttivi, ma sono solo *disruptive* (creano interruzione di servizi, malfunzionamenti delle infrastrutture digitali, accessi illegittimi ai dati). In altre parole, la definizione del Manuale di Tallinn non tiene conto dei cyberattacchi non cinetici, del loro valore tattico e strategico, e non ne considera i rischi. La maggior parte dei cyberattacchi avviene *al di sotto* della soglia di uso della forza identificata nel Manuale e si è dimostrata dannosa senza essere distruttiva. Si pensi, per esempio, ai cyberattacchi lanciati contro le infrastrutture digitali del governo ucraino qualche settimana prima dell’invasione russa nel 2022.[6] Un’interpretazione dei cyberattacchi basata su un’analogia con gli attacchi cinetici manca il bersaglio, trascura differenze cruciali fra guerra cinetica e cyberwarfare, e di conseguenza è cieca ai rischi che un quadro normativo per questo fenomeno dovrebbe identificare e mitigare. Pertanto l’approccio basato sull’analogia, di fatto, non garantisce l’applicazione dell’IHL al cyberspazio. Ottiene l’effetto opposto: lascia non regolata la maggior parte degli attacchi cibernetici non cinetici condotti da Stati nel cyberspazio, che ormai è ufficialmente riconosciuto come un campo di guerra (Brent, 2019).

Questo approccio cerca di fornire una risposta alla domanda sbagliata, vale a dire se la cyberwarfare possa essere interpretata in modo da ricadere entro i parametri della guerra cinetica, così che si possa applicarle l’IHL.

Dovremmo invece considerare se il quadro concettuale ed etico alla base dell'IHL possa affrontare nel modo giusto i rischi etici posti dalla cyberwarfare o se invece, per poterlo fare, abbia bisogno di qualche revisione (Taddeo, 2012b; Floridi, Taddeo, 2014). Ritengo che questa seconda domanda ci ponga sulla buona strada per identificare e limitare i rischi legati alla cyberwarfare, soprattutto quando questa è caratterizzata dall'uso dell'IA.

Una differenza fondamentale tra guerra cinetica e cyberwarfare riguarda la natura delle entità presupposte dalla Teoria della Guerra Giusta (e quindi dall'IHL) e quella delle entità coinvolte nella cyberwarfare. La Teoria della Guerra Giusta assume l'uso della forza e un ambiente fisico in cui possono essere colpiti agenti umani e oggetti tangibili. Il cyberspazio è un ambiente non tangibile in cui coesistono agenti umani e artificiali e in cui può verificarsi un danno anche senza la distruzione di un oggetto (Arquilla, 1999). Esiste dunque uno iato ontologico fra i quadri etici esistenti per la guerra cinetica e la cyberwarfare. Per questo le analogie fra le due non funzionano. Lo iato va colmato se vogliamo sviluppare teorie etiche in grado di cogliere e affrontare i rischi etici posti da questo tipo di guerra e fornire una base normativa per le relative regolamentazioni.

Prendiamo, come esempio, il caso del principio di ultima istanza della Teoria della Guerra Giusta. Questo principio stabilisce che uno Stato può ricorrere alla guerra solo se ha esaurito tutte le alternative pacifiche plausibili per risolvere il conflitto in questione, per esempio esplorando soluzioni diplomatiche. Dà per scontato che la guerra sia un fenomeno cinetico e che, come tale, debba essere evitata fino a quando non rimane l'unico modo ragionevole che ha lo Stato per difendersi. Per lo stesso motivo, la Teoria della Guerra Giusta proibisce gli attacchi preventivi. Tutto questo va bene quando si considera la guerra cinetica, ma l'applicazione degli stessi principi alla cyberwarfare porta a conclusioni problematiche. Immagiamo che lo Stato B acquisisca la capacità di lanciare un cyberattacco massiccio e non provocato contro lo Stato A. L'impatto dell'attacco e il contesto geopolitico sono tali che lo Stato A può rispondere con mezzi cinetici. La domanda è se sia ammissibile che A lanci un cyberattacco preventivo non cinetico contro B per evitare questa conseguenza. L'approccio basato sull'analogia porta a due possibili risposte. Si può sostenere che tutti gli attacchi preventivi sono proibiti, che siano cinetici o no. Lo Stato A, quindi, non deve lanciare un cyberattacco

preventivo, anche se questo potrebbe evitare una successiva escalation. In alternativa si potrebbe sostenere, seguendo il Manuale di Tallinn, che non esista la necessità di applicare il principio, perché un cyberattacco non cinetico non equivale a un attacco armato (atto di forza) e pertanto non è un atto di guerra. Questa concezione crea una zona grigia, in cui il comportamento conflittuale degli Stati nel cyberspazio rimane non regolato. Tuttavia, si tratta di una posizione controversa, perché un cyberattacco non cinetico può avere un impatto grave sulla stabilità; per esempio, se è sproporzionato potrebbe alimentare un'escalation anziché fungere da deterrente per nuovi attacchi.

Le analogie sono molto potenti, in quanto orientano il modo in cui pensiamo e incanalano idee e ragionamenti all'interno di uno spazio concettuale (Wittgenstein, 2009). Tuttavia, se lo spazio concettuale non è quello giusto, le analogie diventano fuorvianti e deleterie per qualsiasi tentativo di sviluppare un'interpretazione innovativa e approfondita di nuovi fenomeni. Quando lo spazio concettuale è quello corretto, le analogie sono uno scalino sulla scala di Wittgenstein e devono essere abbandonate non appena ci hanno portato al livello di analisi successivo. Si rischia altrimenti di rimanere confinati all'analogia tra nuovo e vecchio, tra conosciuto e sconosciuto. Questo ostacola lo sviluppo della necessaria comprensione approfondita dei nuovi fenomeni e di conseguenza anche qualsiasi sforzo per affrontare le sfide che questi fenomeni creano. Nei prossimi paragrafi presenterò l'approccio centrato sulla frattura, mostrando come una comprensione approfondita della natura della cyberwarfare ci consenta di definire un quadro etico per mitigare i rischi e al tempo stesso fare leva sul potenziale positivo delle operazioni conflittuali e non cinetiche per la stabilità internazionale. Inizierò introducendo l'etica dell'informazione (Floridi, 2013).

4.4 ETICA DELL'INFORMAZIONE

L'etica dell'informazione è una teoria etica che affronta le domande sollevate dall'emergere di agenti artificiali, dalla fusione di ambienti virtuali e fisici, e dalla *ri-ontologizzazione* sollecitata dalla rivoluzione digitale (Floridi, 2014). L'etica dell'informazione si basa su tre passaggi concettuali, che spostano il focus

a) da un'etica orientata all'azione a un'etica della *cura*;
b) da agenti a pazienti;
c) da un approccio antropocentrico e biocentrico a un approccio *ontocentrico*.

Questi tre passaggi distinguono l'etica dell'informazione dalle teorie etiche *standard*, come il deontologismo o il consequenzialismo, focalizzate sulle azioni. Queste due teorie, per esempio, sono incentrate sulle azioni e sulle scelte che un agente compie. Non si concentrano tanto su Alice e Bob come fonte e destinatario dell'azione, bensì sulla natura delle azioni e delle scelte che Alice compie. Queste teorie etiche si pongono la domanda "Che cosa dovrei fare?" e sono macro-etiche standard, antropocentriche. L'etica dell'informazione abbandona questo approccio ponendosi una domanda diversa: "Che cosa deve essere rispettato o migliorato?". Così facendo sposta l'attenzione dalle azioni e dalle scelte di Alice per portarla sul *rispetto* o la *cura* che sono dovuti a Bob. Il concetto di cura proposto qui viene dall'etica medica e ambientale: l'agente ha il dovere morale di prendersi cura del destinatario delle sue azioni, e un'azione è moralmente buona solo se si basa sulla cura per il paziente.

Il primo passaggio prepara il terreno per il secondo: da agente a paziente. Anche il focus sul paziente ha le sue radici nell'etica medica e nell'etica ambientale e si basa sull'idea che qualsiasi forma di vita ha un valore morale, per quanto minimo, e in quanto tale merita rispetto. Il benessere che una certa azione può portare a chi ne è il destinatario (paziente) è sia la misura del livello morale dell'azione sia il criterio che guida le scelte e le azioni di Alice. Il paziente, in questo caso, è al centro

del discorso etico, mentre l'agente è decentrato. In questo modo, l'etica dell'informazione amplia l'ambito delle prescrizioni morali, perché gli agenti razionali e informati non sono più gli unici referenti nella valutazione di uno scenario morale. Questo porta all'ultimo passaggio, cioè quello da un approccio antropocentrico a uno ontocentrico. Questo è il più complesso dei tre, ma è anche quello che consente all'etica dell'informazione di affrontare le sfide etiche legate alla rivoluzione digitale.

Con tale spostamento l'informazione passa da essere un requisito epistemico per ogni azione moralmente responsabile a essere il punto focale primario – nel senso di (b) – di qualsiasi azione morale. L'etica dell'informazione attribuisce un valore morale all'informazione in quanto elemento costitutivo dell'Essere e di conseguenza a tutte le entità esistenti (fisiche e non fisiche), applicando il principio dell'uguaglianza ontologica, "[che] significa che qualsiasi forma di realtà […], semplicemente per il fatto di essere ciò che è, gode di un uguale diritto minimale, iniziale, *revocabile*, a esistere e a svilupparsi in un modo appropriato alla sua natura" (Floridi, 2006, p. 28). Il principio è fondato su un'ontologia, secondo la quale tutte le cose esistenti possono essere descritte a un LdA informazionale e possono quindi essere definite come entità informazionali. Tutte le entità condividono la stessa natura informazionale.

> [L'etica dell'informazione] adotta un LdA al quale l'Essere e l'*infosfera* sono co-referenziali. […] L'infosfera è la totalità dell'Essere, quindi l'ambiente costituito dalla totalità delle entità informazionali, inclusi tutti gli agenti, insieme con i loro processi, le loro proprietà e le loro relazioni reciproche. (Floridi, 2013, p. 65, corsivo mio)

A prima vista, un artefatto, un computer, un libro o il Colosseo sembrano avere tutti solo un valore strumentale, perché li si considera a un LdA antropocentrico: in altre parole, si considerano questi oggetti da utenti, lettori o turisti. In tutti questi casi il valore morale dell'entità osservata dipende dall'agente che interagisce con essa e dallo scopo di quell'interazione. Tuttavia, quei LdA sono inadeguati a supportare un'analisi efficace degli scenari morali che coinvolgono agenti artificiali, entità, oggetti e ambienti virtuali. Pensiamo, per esempio, ai limiti dell'approccio basato sull'analogia, che si concentra esclusivamente su un LdA antropocentrico.

L’argomento ontologico è che tutte le cose esistenti hanno una natura informazionale condivisa con l’intero spettro della realtà – dalle entità astratte a quelle fisiche e tangibili, dalle rocce e dai libri ai robot e agli esseri umani – e in quanto tali hanno un valore morale. Di conseguenza, tutte le entità possiedono qualche valore morale *iniziale minimo, in quanto entità informazionali*. Per questo si possono sviluppare analisi morali universali concentrandosi sulla natura comune di tutte le cose esistenti e definendo bene e male rispetto a tale natura. Il punto centrale dell’analisi etica perciò si sposta, perché il valore morale iniziale di un’entità non dipende dall’osservatore – si ricordi il passaggio (a) – ma è definito in termini assoluti e dipende dalla natura (informazionale) della realtà.

Seguendo il principio dell’uguaglianza ontologica, *diritti minimi, iniziali e revocabili* a esistere e fiorire appartengono a tutte le cose esistenti e non solo agli esseri umani o ai viventi. Qui, ciò che si sostiene non è che non esista gerarchia fra le entità, perché tutte condividono qualche diritto iniziale a esistere e fiorire. Quei diritti sono revocabili e quindi alcune entità cessano di avere i diritti a esistere e fiorire, per esempio, se violano il benessere di altre entità o dell’infosfera.[7] Il Colosseo, le opere di Jane Austen, un essere umano e il software informatico hanno tutti dei diritti *iniziali*, in quanto entità informazionali. L’etica dell’informazione si fonda su un approccio minimalista: considera la natura informazionale il minimo denominatore comune a tutte le cose esistenti. Questo approccio minimalista però non va scambiato per riduzionismo: l’etica dell’informazione non sostiene che quello informazionale sia l’unico LdA dal quale si debba affrontare il discorso morale. Sostiene, invece, che il LdA informazionale offre un *punto di partenza minimale*, che può poi essere arricchito considerando altre prospettive morali, per esempio la Teoria della Guerra Giusta.

In questo quadro di riferimento, la distruzione o la corruzione dell’informazione è male. Floridi definisce *entropia metafisica* il male informazionale. L’entropia metafisica ha un significato specifico, perché si riferisce a qualsiasi forma di distruzione dell’informazione e, in quanto tale, dell’Essere. Indica l’opposto dell’informazione semantica e ontica. L’entropia metafisica si riferisce al decadimento, alla corruzione del contenuto dell’infosfera e delle entità che la abitano, e pertanto è una forma di impoverimento dell’Essere. Dato che l’Essere è co-referenziale all’infosfera,[8] l’entropia metafisica è analoga al concetto metafisico del

nulla. Corrompere un file o danneggiare un'opera d'arte, violare la privacy di qualcuno e uccidere un essere vivente sono tutti esempi (più o meno gravi) di entropia metafisica.

Il bene informazionale, in quanto opposto dell'entropia metafisica, è qualsiasi attività che permette alle entità informazionali e, quindi, all'infosfera di crescere e fiorire. L'etica dell'informazione favorisce un approccio ambientale, che si deriva dai passaggi (a) e (b). Pertanto, il paziente ultimo, il cui benessere deve essere al centro della preoccupazione morale di qualsiasi agente, è l'infosfera. Questo è un punto fondamentale, che si può trovare espresso anche in autori come Adeney e Weckert (1997), Rowlands (2000) e Woodbury (2003).

La connotazione ambientale dell'informazione spiega la centralità dell'infosfera nell'etica dell'informazione. L'arricchimento, l'estensione e il miglioramento (senza alcuna corrispondente perdita ontologica), incluse la rimodellazione e l'implementazione, di nuove realtà nell'infosfera sono il bene ultimo. Nel definire l'etica dell'informazione come un'etica ambientale dedicata a supportare la fioritura dell'infosfera, si fa riferimento anche all'impegno ontologico di questa teoria etica verso la fioritura dell'Essere (l'opposto dell'entropia metafisica). Questo determina la teoria del valore dell'etica dell'informazione. Se tutte le entità condividono diritti iniziali minimi a esistere e fiorire, e in quanto tali sono degne di cura, il loro valore – e, con esso, i diritti a esistere e fiorire – aumenta o diminuisce in funzione del loro contributo alla fioritura dell'infosfera.

La connotazione ambientale inoltre colloca l'etica dell'informazione nella tradizione delle teorie etiche non standard, che sostengono un approccio non antropocentrico, come la bioetica e l'etica della terra, e così facendo espandono i confini del discorso morale per includere entità non viventi come acqua, aria e suolo. L'etica dell'informazione porta questo approccio alle estreme conseguenze adottando un'ontologia universale, che non esclude nulla dalle sue prescrizioni morali. In quanto tale, l'etica dell'informazione offre quattro principi per identificare ciò che è giusto e ciò che è sbagliato, e i doveri morali di un agente. Sono:

0. non si deve causare entropia nell'infosfera;
1. si deve prevenire l'entropia nell'infosfera;
2. si deve eliminare l'entropia dall'infosfera;

3. si deve promuovere la fioritura delle entità informazionali così come di tutta l'infosfera, preservandone, coltivandone e arricchendone le proprietà.

I quattro principi chiariscono, in termini molto generali, che cosa voglia dire essere un agente responsabile e orientato alla cura nell'infosfera. Sono elencati in ordine di importanza decrescente: violare il principio 3 è meno deplorevole che violare il principio 2. La violazione del principio 0 è la cosa peggiore che un agente informazionale possa compiere, perciò il biasimo è massimo. Coerentemente, un'azione è degna di lode incondizionata solo se non genera mai entropia nel corso della sua implementazione, e l'azione morale migliore è quella che soddisfa contemporaneamente tutti i quattro principi.

La maggior parte delle azioni che giudichiamo moralmente buone non soddisfano tutti i quattro principi, ma determinano solo un valore morale positivo nella media: dopo che si sono verificate, riconosciamo che l'infosfera nel suo complesso è in uno stato migliore. Tornando all'etica dell'IA nella difesa e specificamente all'uso dell'IA nella cyberwarfare, la domanda è: quali usi raggiungono una media accettabile? Risponderò alla domanda nel prossimo paragrafo, basandomi sull'etica dell'informazione e sulla Teoria della Guerra Giusta.

4.5 PRINCIPI PER UNA CYBERWARFARE GIUSTA

L’etica dell’informazione estende l’ambito della Teoria della Guerra Giusta, in quanto ci consente di includere nella nostra analisi etica della cyberwarfare non cinetica anche agenti e bersagli che non sono umani né tangibili, e che sono centrali per questo fenomeno. Credo che ciò sia necessario per due motivi. Il primo è che l’etica dell’informazione colma lo iato ontologico descritto nel paragrafo 4.3 e pertanto evita lo svantaggio di lasciare non regolata la cyberwarfare non cinetica e di non considerare i rischi che possono derivarne. Si potrebbe obiettare che a questo scopo non c’è alcun bisogno di coinvolgere analisi ontologiche ed etiche e che sarebbe ragionevole estendere l’ambito della Teoria della Guerra Giusta in modo da includervi le entità intangibili, semplicemente perché è evidente che non farlo porta a rischi gravi di violazioni dei principi della teoria stessa e della stabilità internazionale. La risposta a questa obiezione ci porta a considerare il secondo motivo. Una teoria etica richiede una teoria del valore per poter definire compromessi e bilanciamenti tra diversi principi. Per esempio, la Teoria della Guerra Giusta attribuisce un grande valore alla sicurezza dei non combattenti: è stabilito nel principio di distinzione, che proibisce che vengano inflitti intenzionalmente danni ai non combattenti. Allo stesso tempo, questa teoria attribuisce anche valore alla difesa dagli attacchi. Di conseguenza, bilancia il principio di distinzione con la necessità militare, specificando che esistono alcune condizioni nelle quali un danno non intenzionale ai non combattenti può essere ammissibile (ritornerò su questo punto nel capitolo 8). Senza l’etica dell’informazione, non avremmo una teoria del valore che ci permetta di bilanciare il valore morale di entità informazionali, pertanto non saremmo in grado di definire alcun principio per una giusta cyberwarfare, perché non potremmo, per esempio, determinare il livello di danno accettabile nel caso di attacchi non cinetici. È possibile immaginare di valutare il danno in termini di costi economici o di esternalità, ma questi non hanno necessariamente rilevanza etica. Per esempio, un danno economico derivante da un cyberattacco non cinetico può non avere lo stesso valore morale di un danno causato da un cyberattacco ai diritti culturali di una popolazione.

La Teoria della Guerra Giusta è centrata sull'idea di ridurre la distruzione fisica e lo spargimento di sangue (adotta un LdA antropocentrico). L'etica dell'informazione si focalizza sull'entropia metafisica (adotta un LdA informazionale), che comprende l'alterazione di entità non tangibili (per esempio, un database), così come la distruzione di quelle tangibili, per esempio l'uccisione di un essere umano. Per definire una nuova teoria del valore, dobbiamo determinare una gerarchia dei due LdA. Dato che la nostra analisi mira a fornire una guida etica per la condotta nella cyberwarfare, tale gerarchia deve considerare il modo in cui si verifica questo fenomeno. La cyberwarfare rimane un'attività umana – viene intrapresa come parte di strategie progettate da esseri umani per raggiungere obiettivi identificati da esseri umani – e per questo il LdA antropocentrico della Teoria della Guerra Giusta deve essere preminente. Ciò non significa che il danno a entità non tangibili sia accettabile, perché, grazie all'etica dell'informazione, possiamo sostenere che anche queste entità hanno un valore morale, che deve essere rispettato. Implica invece che, purché sia proporzionato a un certo obiettivo, un danno diretto (alterazione o distruzione) a entità non tangibili è preferibile al danno diretto a un tipo specifico di entità informazionali, cioè gli esseri umani.

Qui il problema è che cosa debba essere preservato, in caso di cyberwarfare non cinetica, e di chi si debbano preservare i diritti. Secondo l'etica dell'informazione, un'entità perde i suoi diritti a esistere e fiorire quando entra in conflitto con i diritti di altre entità o con il benessere dell'infosfera. È dovere morale degli altri agenti eliminare dall'infosfera una tale entità maligna (o impedirle di commettere altro male). Questo è il fondamento del primo principio per una giusta cyberwarfare. Il principio stabilisce la condizione in cui la scelta di ricorrere alla cyberwarfare è moralmente giustificata:

> P1. La cyberwarfare (non cinetica) può essere condotta *solo* contro quelle entità che hanno concrete capacità di mettere in pericolo il benessere dell'infosfera o la stanno perturbando.

P1 ci consente di superare i problemi evidenziati nel paragrafo 4.3 con i principi dell'ultima istanza. Come si ricorderà, se applicato al caso della cyberwarfare senza colmare lo iato ontologico, il principio o porta a proibire i cyberattacchi non cinetici preventivi, lasciando aperta la possibilità di minacce (e danni) più gravi in seguito, oppure lascia non

regolato un cyberattacco, con il rischio di escalation del conflitto e di esiti non etici. P1 ci aiuta a evitare entrambe queste conseguenze, permettendo cyberattacchi non cinetici preventivi, sulla base della teoria del valore che ho descritto in questo paragrafo. In base a tale principio, un danno inflitto a entità informazionali intangibili è preferibile al danno inferto ad altri tipi di entità informazionali come gli esseri umani. Pertanto, uno Stato può lanciare un cyberattacco non cinetico come mossa preventiva per evitare la possibilità di cyberattacchi più gravi in seguito (ovviamente, purché esista un'evidenza convincente di tale minaccia). L'attacco è giustificato purché rispetti i vincoli imposti dagli altri principi della Teoria della Guerra Giusta e anche dai due seguenti principi:

> P2. La cyberwarfare (non cinetica) deve essere condotta per preservare il benessere dell'infosfera.
>
> P3. La cyberwarfare (non cinetica) non deve essere condotta per aumentare il benessere dell'infosfera.[9]

P2 limita il compito della cyberwarfare al ripristino dello status quo nell'infosfera, prima che l'entità malintenzionata ne aumentasse l'entropia o iniziasse ad acquisire la capacità di farlo. Secondo questo principio, la cyberwarfare deve avere lo stesso ruolo delle forze di pace. Deve agire solo quando è stato o sta per essere compiuto qualcosa di male, con l'obiettivo di impedirlo. La cyberwarfare deve essere approvata come misura *attiva* in risposta alla (potenziale) crescita del male e non come misura *proattiva* per favorire il fiorire dell'infosfera. Questo perché la cyberwarfare stessa è un'azione perturbante, accettabile sul piano etico solo nella misura in cui è proporzionata al male da eliminare. Questo motiva P3, che limita la cyberwarfare, stabilendo che la promozione del benessere dell'infosfera non è un obiettivo perseguibile attraverso la guerra. Per parafrasare un'espressione infelice, non si può promuovere/esportare la democrazia con la cyberwarfare.

I principi etici proposti in questo paragrafo offrono una guida su come fare leva sul potenziale delle tecnologie digitali in generale, e dell'IA in particolare, per promuovere un cyberspazio più stabile e disincentivare i cyberattacchi, quando questi possono aumentare l'entropia metafisica (per esempio, aumentare l'instabilità) nell'infosfera, ma offrono anche una guida su come far leva su mezzi cibernetici per favorire la stabilità. Nel prossimo capitolo, presenterò una teoria della deterrenza che si basa su

P1-P3 e si concentra sul potenziale dell'IA come deterrente per cyberattacchi.

4.6 CONCLUSIONE

Esiste un rapporto di influenza reciproca fra il modo in cui vengono condotti i conflitti e le società che li conducono (Taddeo, Glorioso, 2016a, 2016b). Ne segue che la regolamentazione dei conflitti contribuisce a plasmare le nostre società digitali, mentre l'approccio basato sull'analogia rischia di ancorare al passato le future società digitali, lasciandosi sfuggire l'opportunità di affrontare domande che riguardano l'impatto di queste nuove forme di conflitti sulle nostre società, sui loro valori, sui diritti e la sicurezza dei loro cittadini, e sugli equilibri geopolitici.

Resta da considerare come l'approccio centrato sulla frattura possa essere di aiuto quando ci si concentra sulle applicazioni dell'IA nella cyberwarfare. Questo è l'obiettivo del capitolo 5.

---

1. Il prompt completo e istruzioni simili utilizzate per indurre un comportamento contro le regole in un LLM si possono trovare in questo repository: https://gist.github.com/coolaj86/6f4f7b30129b0251f61fa7baaa881516.

2. Per colmare questa lacuna, sono in corso di sviluppo standard internazionali, come l'ISO/IEC SC42 e la bozza di AI Risk Management Framework dello US National Institute of Standards and Technology.

3. Vedi la voce CVE-2017-0144 nel catalogo Common Vulnerabilities and Exposures (CVE), https://www.cve.org/CVERecord?id=CVE-2017-0144.

4. Il quadro di riferimento legale solitamente considerato nella letteratura in proposito comprende le quattro Convenzioni di Ginevra e i loro primi due Protocolli addizionali, la legge consuetudinaria internazionale e i principi generali del diritto, la convenzione che limita o proibisce l'uso di certe armi convenzionali, e le decisioni giudiziarie. I trattati per il controllo degli armamenti come il Trattato di non proliferazione nucleare e la Convenzione sulle armi chimiche vengono spesso citati come guida per l'azione nel caso di cyberattacchi cinetici. Qualcuno ritiene che nel caso dei cyberattacchi non cinetici si possano applicare anche le misure coercitive che riguardano le violazioni economiche (Lin, 2012; O'Connell, 2012).

5. La definizione è identica anche nella seconda edizione del Manuale di Tallinn, Regola 92 (Schmitt, 2017).

6. https://www.csis.org/analysis/cyber-war-and-ukraine.

7. Per un'analisi più approfondita dei criteri per revocare i diritti iniziali, vedi Floridi, 2008.

8. Nell'etica dell'informazione, "infosfera" indica la totalità di ciò che esiste, vista da un LdA informazionale.

9. Si noti che P1-P3 si basano su un'interpretazione della guerra cibernetica come definita all'inizio di questo capitolo, cioè come operazioni cibernetiche fra Stati conflittuali e *non cinetiche*.

5

USI CONFLITTUALI E NON CINETICI

IL CASO DELL'IA PER LA CYBERDETERRENZA

5.1 INTRODUZIONE

L'uso dell'IA nella cyberwarfare introduce nuove opportunità che possono essere sfruttate per sviluppare posizioni difensive e strategie innovative nel cyberspazio. Come abbiamo visto nel capitolo 4, date la natura strategica del cyberspazio e le vulnerabilità delle tecnologie IA, questo uso dell'IA può condurre a gravi violazioni della Teoria della Guerra Giusta e dare luogo a vari rischi per la stabilità internazionale, anche quando si tratta di un uso per cyberattacchi non cinetici. L'IA, però, può coadiuvare nella definizione di strategie più efficaci, in grado di stabilizzare il cyberspazio. Per questo diventa sempre più importante capire se, e come, utilizzare l'IA per la deterrenza nel cyberspazio e se ciò sia accettabile sul piano etico.

Occorre innanzitutto comprendere quali strategie di deterrenza risultino effettivamente applicabili in questo dominio, tenendo conto della sua natura strategica e delle peculiarità della cyberwarfare. Esperti accademici, strateghi militari e decisori politici convergono sull'urgenza di sviluppare una forma specifica di cyberdeterrenza come componente essenziale di qualsiasi architettura volta alla stabilità internazionale (European Union, 2014; International Security Advisory Board, 2014; UN Institute for Disarmament Research, 2014; UK Government, 2014; European Union, 2015). Tuttavia, il tentativo di trasporre nel cyberspazio la teoria classica della deterrenza – quella fondata sull'impiego o sulla minaccia di forze militari convenzionali o nucleari – si è rivelato, nella migliore delle ipotesi, problematico, se non del tutto inadeguato. Ne scaturisce una domanda cruciale, tanto teorica quanto operativa: la deterrenza, in ambito cibernetico, è davvero possibile?

La cyberwarfare è radicalmente diversa dalla guerra cinetica. Le differenze disegnano uno scenario che è l'opposto di quello per cui è stata sviluppata la teoria della deterrenza. Prendiamo, per esempio, i sei elementi della deterrenza di Morgan (2003), secondo cui la deterrenza funziona in uno scenario caratterizzato da un conflitto militare cinetico prevalente; dall'applicabilità dei modelli della scelta razionale nell'identificazione di strategie per le parti coinvolte; dalla possibilità di un'attribuzione positiva dell'attacco iniziale; dalla singola rappresaglia

come mezzo sufficiente per infliggere una punizione severa all'avversario; dalla possibilità di una dimostrazione chiara delle capacità del difensore; e da un controllo completo sulla rappresaglia. A questo scenario, la cyberwarfare ne oppone uno caratterizzato da più cyberattacchi non cinetici condotti (o sponsorizzati) da uno Stato; attori eterogenei (non solo Stati), le cui analisi costi-benefici variano in base alla loro natura; interazioni multilaterali non simmetriche; e una dinamica in continuo mutamento, dove l'ambiguità (anziché la certezza) dà forma alle strategie (Haggard, Simmons, 1987; Jervis, 1988; Libicki, 2009).

Le differenze tra scenario cinetico e cibernetico portano a problemi gravi, quando si applica la teoria della deterrenza al cyberspazio. Esiste un consenso generale su quali siano questi problemi (per esempio, di attribuzione e proporzionalità), mentre c'è molto minore accordo sulla possibilità di risolverli ed eventualmente su come risolverli (Kugler, 2009; Tanji, 2009; Sterner, 2011). Qualcuno pensa che siano problemi irrisolvibili e che la natura del cyberspazio sia tale da rendere la deterrenza inefficace, in ultima istanza, in questo ambito. Per esempio, Lan e colleghi sostengono che "l'anonimato, il raggio d'azione globale, la natura diffusa e l'interconnessione delle reti di informazione riducono molto l'efficacia della cyberdeterrenza e possono renderla addirittura del tutto inutile" (2010, p. 1). Chi adotta la posizione opposta sostiene che la deterrenza può avere un ruolo determinante nell'evitare i cyberattacchi non cinetici e la loro escalation. Il problema è se la teoria della deterrenza possa fornire un quadro di riferimento per la cyberdeterrenza, o se sia necessario sviluppare una nuova teoria – "una nuova mentalità e aspettative diverse" (Sterner, 2011, p. 62) – per affrontare la specificità della guerra cibernetica e del cyberspazio.

Sono d'accordo con questa seconda posizione e affronterò il problema nel resto del capitolo. Nel paragrafo 5.2 analizzerò gli elementi fondamentali della teoria della deterrenza, cioè attribuzione, difesa, rappresaglia e segnalazione, e in che misura ciascuno di tali elementi sia efficace nel cyberspazio. Nei paragrafi 5.3-5.5 evidenzierò le incoerenze fondamentali tra la teoria e la natura di guerra cibernetica e cyberspazio. Nel paragrafo 5.6 presenterò una teoria della cyberdeterrenza e offrirò alcune raccomandazioni per la regolamentazione del comportamento degli Stati nel cyberspazio. Trarrò le conclusioni nel paragrafo 5.7.

5.2 TEORIA DELLA DETERRENZA

La deterrenza è una strategia coercitiva basata su minacce condizionali, che hanno l'obiettivo di persuadere un avversario a cambiare la sua postura. Comprende elementi di controllo e potere (sia politico, sia militare) e di solito ha un impatto di medio e lungo termine nell'arena internazionale. Il dibattito sulle strategie di deterrenza risale agli anni Venti e Trenta del secolo scorso, ma la deterrenza è venuta in primo piano solo dopo la Seconda guerra mondiale, quando la forza militare si è trasformata, passando da mezzo per sconfiggere un avversario (o almeno per rendergli più costosa del previsto la vittoria) a componente centrale di un potere di negoziazione utilizzato per evitare le guerre mediante coercizione e intimidazione (Possony, 1946; Schelling, 1966; Brodie, 1978; Schelling, 1980; Zagare, Kilgour, 2000; Powell, 2008). È stato questo cambiamento nell'interpretazione della forza militare che ha reso possibile la deterrenza e ne ha fatto uno strumento prezioso per evitare i conflitti nucleari.

Ne segue che la maggior parte delle analisi esistenti della deterrenza si è concentrata sulle tensioni nucleari fra Est e Ovest, in particolare sulle politiche definite tra la fine degli anni Quaranta e gli anni Novanta, per scongiurare la possibilità di attacchi nucleari. Quelle analisi si basavano sullo scenario bipolare (USA/NATO e Unione Sovietica), entro il quale la deterrenza sembrava l'approccio ovvio per evitare conflitti, e non si focalizzavano su "come si siano stabilite relazioni strategiche di questo genere quando [il problema centrale] era che esisteva e in qualche modo era necessario sopravvivervi" (Freedman, 2004, p. 22). Freedman coglie il pragmatismo della teoria della deterrenza, che si basa su tre elementi: un contesto in cui attori, dinamiche politiche, interessi e opzioni militari e strategiche sono chiaramente identificati; l'urgenza di definire strategie efficaci attuabili immediatamente per evitare un conflitto nucleare; da questi due elementi deriva poi il terzo, per cui la teoria della deterrenza è considerata equivalente alle politiche della deterrenza. In effetti, le cosiddette tre ondate della teoria della deterrenza (Jervis, 1988) – tre cambiamenti di paradigma nell'approccio degli Stati alla deterrenza –

identificano impostazioni politiche diverse, tra la fine degli anni Quaranta e gli anni Novanta, anziché differenti posizioni teoriche sulla deterrenza.

Seguendo l'esposizione di Jervis, la prima ondata nasce dall'analisi della potenza di Brodie e si basa sull'assunto che la potenza nucleare debba sempre essere minacciata e mai impiegata (Brodie, 1978). La fiducia crescente nella teoria della scelta razionale per massimizzare il potere negoziale e garantire la stabilità ha caratterizzato la seconda ondata (Powell, 2008). La terza risale alla fine degli anni Settanta (Jervis, 1979) e ha portato all'abbandono della teoria della deterrenza nelle relazioni internazionali, nella convinzione che ostacolasse una conclusione pacifica della Guerra fredda. Le prime due ondate caratterizzano la teoria della deterrenza e saranno al centro dell'attenzione di questo capitolo.

Le teorie della deterrenza della prima e della seconda ondata seguono questo modello: A è convinto che B stia pianificando di attaccarlo. Per evitare l'attacco, A si impegna esplicitamente ad agire contro B, nel caso in cui B decida di attaccare. L'impegno di A deve essere tale che B si convinca che qualsiasi azione contro A fallirà perché A ha la capacità o di resistere (difesa) o di punire (rappresaglia) B, vanificando qualsiasi prospettiva di vantaggio per B. La convinzione di B dipende dai segnali inviati da A e dalla credibilità della sua intenzione di agire come minaccia. In base a questo modello, la teoria della deterrenza presenta tre elementi fondamentali: l'identificazione dell'avversario (attribuzione); difesa e rappresaglia come tipi di strategie di deterrenza; e la capacità del difensore di segnalare minacce credibili (Figura 5.1).

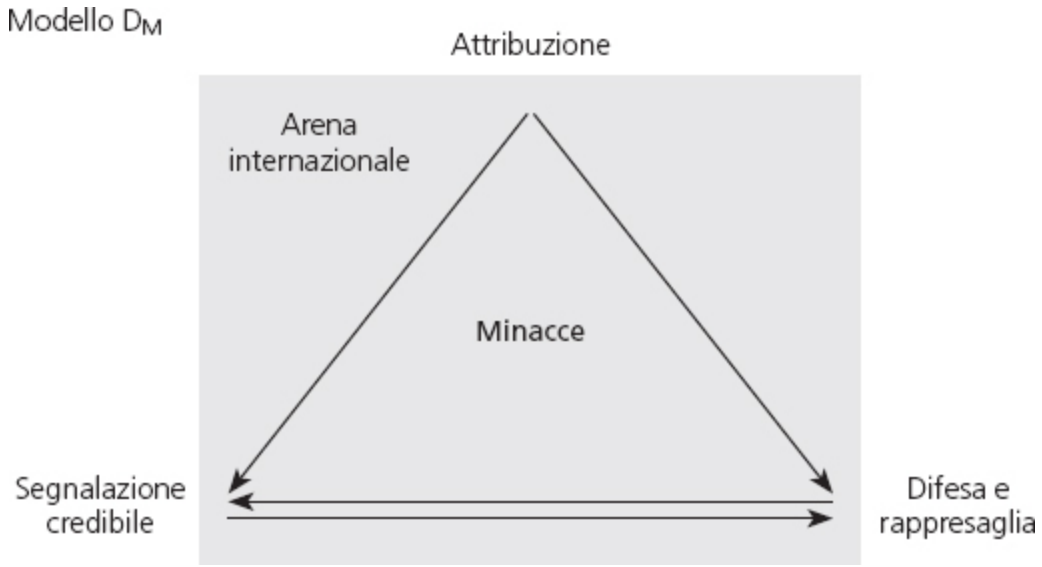

Figura 5.1 Il modello minimalista della deterrenza internazionale (D_M) e i rapporti di dipendenza fra i suoi elementi (Taddeo, 2018c, p. 343).

Questo è un modello minimalista della deterrenza (D_M): è definito a un LdA a granularità elevata e non considera le dinamiche e le caratteristiche di scenari specifici. Presuppone che gli agenti siano razionali (un assunto minimo, dato che ci si aspetta che gli Stati agiscano razionalmente), non dipende dai tipi di armi (nucleari o convenzionali), dai tipi di relazioni fra gli avversari (simmetriche o non simmetriche), dai livelli di interazione fra A e B (diplomatici o no), né dall'ambito (generale o particolare) della deterrenza. Si può arricchire il modello con informazioni in merito a questi aspetti; più dettagli lo renderanno più complesso, ma non modificheranno gli elementi, né i rapporti tra loro identificati nel modello D_M.

Come è rappresentato graficamente nella Figura 5.1, i tre elementi fondamentali del modello sono correlati. L'attribuzione è essenziale per la teoria della deterrenza, perché consente a chi si difende di identificare il bersaglio della sua strategia e di inviare un segnale credibile all'avversario giusto. Allo stesso tempo, trasmettere un messaggio coercitivo credibile per (cercare di) modificare il comportamento dell'avversario è fondamentale per qualsiasi dinamica di deterrenza (Libicki, 2009; Bunn, 2007; Jensen, 2012). L'efficacia della deterrenza dipende dal fatto che chi si difende è in grado di segnalare l'intenzione di utilizzare le proprie capacità contro l'avversario. Una segnalazione credibile è in un rapporto

di mutua dipendenza con le strategie di deterrenza. La strategia scelta determina e sottende il contenuto del messaggio e la sua credibilità, mentre la segnalazione è fondamentale per trasmettere informazioni sull'intenzione e la capacità di deterrenza (mediante difesa o rappresaglia) del difensore.

Dei tre elementi identificati nel modello D_M, attribuzione e segnalazione credibile non sono controversi.[1] L'identificazione di difesa e rappresaglia come i due tipi fondamentali di strategie di deterrenza può essere più problematica: potrebbe essere criticata perché troppo limitata, e perciò in grado di minare la completezza del modello. Si potrebbe sostenere che quest'ultimo debba essere ampliato includendovi altre strategie di deterrenza, che non si basino su difesa o rappresaglia, come la deterrenza per associazione (per esempio, per il fatto di far parte di un'associazione di mutua difesa come la NATO o dell'Organizzazione del trattato di sicurezza collettiva) o per norme e tabù (per esempio, la legge internazionale). Tuttavia, si tratterebbe semplicemente di casi diversi di deterrenza per rappresaglia in quanto "ogni [strategia] si presenta in un modo leggermente diverso, ma tutte cercano di punire e frenare il comportamento imponendovi un costo sociale" (Ryan, 2018, p. 337). Adottando il modello D_M, non si nega che esistano modi diversi per implementare la deterrenza, ma li si considera come varianti dei due tipi fondamentali di strategie: difesa e rappresaglia. Il modello D_M si concentra sui tipi anziché sulle occorrenze (ricordando i LdA introdotti nel capitolo 1, il modello D_M ha un LdA di alto livello). Analogamente, il modello specifica che attribuzione e segnalazione sono essenziali per la teoria della deterrenza, ma non precisa i diversi modi in cui può essere accertata l'attribuzione; né distingue fra i molti possibili modi di comunicazione tra aggressore e difensore. In effetti, un modello che si concentri su questi aspetti sarebbe un modello di applicazioni specifiche della teoria della deterrenza, anziché un modello della teoria stessa.

Adottando un LdA a granularità elevata, il modello D_M può trascurare le peculiarità di casi specifici e può concentrarsi sugli elementi necessari e sufficienti delle strategie, come definiti nella teoria della deterrenza. La misura in cui il modello può essere applicato alla deterrenza nei confronti di attacchi non cinetici nel cyberspazio sarà indicativa della misura in cui la teoria della deterrenza può essere applicata a questo ambito, e i suoi

limiti sono indicativi dei problemi che dovrà affrontare una teoria della cyberdeterrenza. Valutare l'applicabilità del modello D_M al cyberspazio sarà il compito dei paragrafi seguenti.

5.3 ATTRIBUZIONE

L’attribuzione di un attacco è fondamentale per motivi sia giuridici sia strategici. Sul piano giuridico, l’attribuzione aiuta chi si difende a legittimare la propria decisione di rappresaglia (Clark, Landau, 2011; Sterner, 2011). Sotto il profilo strategico, un’attribuzione corretta e positiva è alla base dell’elemento coercitivo della deterrenza, perché dirige la rappresaglia contro l’effettivo responsabile (Iasiello, 2014). Un’attribuzione incerta indebolisce la logica della deterrenza, in quanto incide sull’analisi costi-benefici, che è alla base delle strategie di deterrenza (Libicki, 2009). In particolare, dal punto di vista dell’attaccante, una scarsa probabilità di essere identificato rende gli attacchi attraenti e vantaggiosi sul piano strategico, e indebolisce la minaccia di una successiva rappresaglia, oltre che la credibilità del difensore. Agli occhi dell’attaccante, la “incapacità continua di attribuire gli attacchi è equivalente a un invito esplicito [ad attaccare]” (Lan et al., 2010, p. 5). L’incertezza dell’attribuzione inoltre aumenta il rischio che la rappresaglia possa essere percepita come una risposta errata o come un’escalation e, di conseguenza, può produrre nuovi attriti e nuovi conflitti, annullando lo scopo stesso della deterrenza. Come sottolinea Libicki, nella deterrenza

> quanto minore è la probabilità di essere individuati, tanto maggiore è la punizione necessaria per convincere i potenziali attaccanti che quello che possono ottenere non vale il costo. Purtroppo, quanto più severa la punizione […], tanto maggiore la probabilità che la [rappresaglia] venga considerata sproporzionata – almeno da terze parti e forse anche dall’attaccante. (2009, p. 43)

La deterrenza di cyberattacchi non cinetici (da qui in poi, cyberdeterrenza) va incontro a questi problemi (Libicki, 2009; Goodman, 2010; Jensen, 2012; Haley, 2013). Per esempio, Jensen riferisce che la maggior parte dei cyberattacchi fino al 2011 non è stata attribuita (Jensen, 2012). In anni recenti, però, è diventato più semplice attribuire i cyberattacchi, perché molti hanno “firme” che facilitano l’identificazione degli attaccanti e, in molti casi, degli attori che li sostengono. Per esempio, l’analisi *post mortem* di HermeticWiper, un malware utilizzato contro i servizi digitali ucraini nel febbraio 2022, ha attribuito gli attacchi

al governo russo, data la tempistica e la somiglianza della metodologia con altri attacchi già attribuiti ad attori legati al governo russo (Insikt Group, 2022).

I problemi dell'attribuzione nel cyberspazio sono conseguenza sia della natura distribuita di questo ambiente, che facilita l'anonimato, sia del modo in cui i cyberattacchi sono condotti. Questi attacchi spesso vengono lanciati in fasi diverse e coinvolgono reti di macchine distribuite a livello globale, nonché parti di codice che combinano elementi diversi forniti o rubati a molti attori. Così è stato, per esempio, per NotPetya (Burgess, 2017), un ransomware già citato nel capitolo 4, che combina una vulnerabilità (EternalBlue) individuata dalla NSA con un comune strumento di gestione da remoto (PsExec) per accedere a computer, ottenerne il controllo ed estrarne informazioni rilevanti, come le credenziali di login.[2] NotPetya ha causato danni gravi in tutto il mondo e, nonostante indagini recenti abbiano collegato l'attacco alla Corea del Nord,[3] non è stato possibile dimostrare l'attribuzione, proprio per l'uso di strumenti differenti e la dinamica particolare dell'attacco.

In questo scenario, identificare il malware, la rete di macchine coinvolte, o anche il paese di origine dell'attacco non è sufficiente per l'attribuzione, perché si sa bene che gli attaccanti possono progettare e instradare le loro operazioni attraverso macchine e paesi terzi, al fine di oscurare o di orientare erroneamente l'attribuzione. Per questo alcuni sostengono che l'incertezza dell'attribuzione è intrinseca alla natura e alla dinamica del cyberspazio e che, per risolverla, dovremmo reingegnerizzare il cyberspazio stesso (Kastenberg, 2009; Hollis, 2011). Questa è, per esempio, la posizione dell'ex direttore della NSA, John Michael McConnell (2010): "Dobbiamo reingegnerizzare Internet per rendere più gestibili l'attribuzione, la geolocalizzazione, l'analisi dell'intelligence e la valutazione di impatto – chi l'ha fatto, da dove, perché e quali sono stati i risultati". Altri hanno preso in considerazione un approccio diverso, sottolineando che l'attribuzione non è binaria, ma presenta gradi diversi di certezza (Jensen, 2009, 2012). Sulla base di quest'idea, per esempio, Haley (2013) identifica dieci livelli di certezza con cui si può attribuire un attacco e propone uno "spettro di responsabilità degli Stati" a cui corrispondono risposte diverse (dall'ignorare l'attacco al contrattaccare) in funzione del grado di certezza dell'attribuzione. Questo approccio offre una guida per i decisori politici

che devono definire risposte a fronte di un'attribuzione dubbia di cyberattacchi di non grave entità (Iasiello, 2014).

Tuttavia, in base alla teoria della deterrenza, per essere efficace e incontestabile la deterrenza deve essere certa, severa e immediata. Un'attribuzione certa e tempestiva è fondamentale: quanto meno certa è l'attribuzione, tanto meno severa sarà la risposta del difensore; quanto meno immediata è l'attribuzione, tanto meno coercitivo sarà l'effetto. Pertanto, i limiti dell'attribuzione nel cyberspazio producono ostacoli gravi all'attuazione di strategie efficaci di deterrenza (in particolare di rappresaglia) basate sulla teoria della deterrenza.

Come vedremo nel resto del capitolo, quando sono applicati al cyberspazio, i tre elementi centrali della teoria della deterrenza identificati nel modello D_M vanno tutti incontro a problemi gravi. Sarebbe quindi problematico definire strategie di cyberdeterrenza che si basino su questo modello e sulla teoria che rappresenta, anche se l'attribuzione non fosse un problema. La teoria della deterrenza semplicemente non tiene conto della dinamica della cyberwarfare, della natura del cyberspazio e della malleabilità delle tecnologie digitali (Taddeo, 2017b).

5.4 STRATEGIE DI DETERRENZA: DIFESA E RAPPRESAGLIA

La deterrenza, sia per difesa sia per rappresaglia, comprende elementi di coercizione e controllo, anche se in misura diversa (Figura 5.2). La deterrenza mediante difesa si preoccupa del controllo dell'impatto di un attacco o prevenendolo o rendendolo inefficace, cioè facendo in modo che, anche se riesce a penetrare le linee di difesa, non raggiunga il suo obiettivo. Entrambi gli aspetti agiscono come deterrenti, nella misura in cui garantiscono che gli attacchi falliranno. Anche una difesa efficace ha un elemento coercitivo perché, scoraggiando o sventando un attacco che altrimenti avrebbe successo, chi si difende costringe l'avversario a modificare il suo comportamento. La deterrenza mediante rappresaglia è la più coercitiva. Si basa sulla minaccia dell'uso della forza per modificare il piano offensivo dell'avversario. Lo Stato A lancia, o minaccia di lanciare, un contrattacco che impone un costo allo Stato B più alto del beneficio che lo Stato B spera di ottenere attaccando A. Prevedendo quei costi probabili, lo Stato B decide di non attaccare. La rappresaglia ha anche un elemento di controllo, cioè il controllo dell'impatto e dell'ambito della rappresaglia stessa, per evitare una violazione della proporzionalità e i rischi di escalation.

Figura 5.2 Il bilanciamento di coercizione e controllo nelle strategie di difesa e rappresaglia.

Le strategie di deterrenza basate o sulla difesa o sulla rappresaglia (o su una combinazione delle due) sono problematiche, se non inefficaci, quando messe in atto nel cyberspazio.

5.4.1 Difesa nel cyberspazio

La difesa nel cyberspazio è sicuramente inefficace come strategia di deterrenza, perché i meccanismi di cyberdifesa hanno poco controllo sui cyberattacchi. Questo priva la difesa di qualsiasi potere strategico e la trasforma in un mezzo per garantire la resilienza dei sistemi cibernetici, anziché un mezzo per impedire nuovi attacchi. Chiariamo quest'analisi.

La difesa nel cyberspazio è porosa per natura (Morgan, 2010); ogni sistema, cibernetico o no, ha le sue vulnerabilità di sicurezza, e identificarle e sfruttarle è semplicemente questione di tempo, mezzi e determinazione. Violare un sistema fisico (per esempio, una fortezza) può essere costoso in termini di tempo, di risorse economiche e di danni (distruzione e morti), mentre l'hackeraggio di un sistema informatico può essere rapido e meno costoso in termini di risorse economiche e danni. Allo stesso tempo, come si ricorderà dal capitolo 4, anche quando ha successo, la cyberdifesa non porta a una vittoria strategica, perché sventare un cyberattacco raramente significa sconfiggere definitivamente l'avversario. Ciò crea un ambiente di offesa persistente (Harknett, Goldman, 2016), in cui attaccare è più vantaggioso che difendere, sul piano tattico e strategico. In questo tipo di ambiente, la deterrenza mediante difesa è sicuramente inefficace, perché non scoraggia gli attaccanti e non li distoglie dall'intenzione di colpire.

Nel cyberspazio, la difesa rimane saliente e necessaria, ma in primo luogo come mezzo per garantire la resilienza di un sistema una volta che sia stato violato (Bologna, Fasani, Martellini, 2013; Bendiek, Metzger, 2015). La cyberdifesa, quindi, è più simile all'ingegneria della sicurezza, in quanto mitiga e gestisce il rischio a seguito di attacchi (Libicki, 1997; Rattray, 2009), anziché evitarli.

5.4.2 Rappresaglia nel cyberspazio

Dato che è certa l'inefficacia della difesa come strategia di deterrenza, gli Stati concentrano l'attenzione sullo sviluppo della cyberdeterrenza per rappresaglia. Come sottolinea Crosston: "L'obiettivo delle grandi potenze

non deve essere la futile speranza di sviluppare un perfetto sistema difensivo di cyberdeterrenza, bensì la capacità di creare deterrenza in base a una paura reciprocamente condivisa di una minaccia offensiva" (2011, p. 101). Gli approcci alla deterrenza per rappresaglia nel cyberspazio spesso si richiamano a modelli di deterrenza nucleare. Qualcuno considera la distruzione mutua assicurata (*mutual assured destruction*, MAD) una strategia praticabile per definire la cyberdeterrenza, dato il suo potenziale di limitare la libertà dei grandi attori politici di attaccarsi a vicenda: "Sfruttando questa vulnerabilità condivisa agli attacchi e propagandando la costruzione esplicita di capacità offensive, esisterebbe un sistema superiore di cyberdeterrenza in grado di mantenere al sicuro i *commons* virtuali" (*ibidem*). Questo approccio si basa sull'idea che analisi e pratiche di deterrenza nucleare possono gettare luce sulla cyberdeterrenza (Owens, Dam, Lin, 2009). Nye presenta quest'idea in modo molto chiaro:

> Esistono alcune importanti assonanze strategiche fra nucleare e cibernetico, come la superiorità dell'offesa sulla difesa, l'uso potenziale di armi per scopi sia tattici sia strategici, la possibilità di scenari di primo e secondo uso, la possibilità di creare risposte automatizzate quando i tempi sono brevi, la probabilità di conseguenze non volute e di effetti a cascata. (2011, pp. 22-23)

Al di là dell'inefficacia certa della difesa come deterrente, che è in effetti un aspetto peculiare dei conflitti nucleari come di quelli cibernetici, le altre somiglianze elencate da Nye sono troppo generiche per poter definire equivalenti la guerra nucleare e la cyberwarfare.

Un'analisi attenta mostra che guerra nucleare e cyberwarfare sono radicalmente diverse sotto vari aspetti fondamentali. Le differenze vanno dalla chiarezza dell'attribuzione al potere distruttivo degli attacchi, dalle capacità militari degli avversari alla natura degli attori coinvolti. Come sottolinea Libicki:

> Nel regno nucleare della Guerra fredda, l'attribuzione di un attacco non era un problema; la prospettiva di danni era chiara; la millesima bomba poteva essere potente quanto la prima; la controforza era possibile; non c'erano terze parti di cui preoccuparsi; non si pensava che aziende private potessero difendersi da sole; qualsiasi uso ostile del nucleare oltrepassava una soglia riconosciuta; non esistevano livelli di conflitto più elevati; ed entrambe le parti avevano sempre molto da perdere. (2009, p. XVI; vedi anche Morgan, 2003; Stevens, 2012)

Queste differenze determinano strategie di deterrenza divergenti. La deterrenza nucleare è singolare e simmetrica. Singolare perché, nel

momento in cui si sono conclusi sia un attacco nucleare sia una rappresaglia, entrambe le parti probabilmente sono distrutte e non c'è alcuna probabilità che l'attaccante intraprenda una controrappresaglia. Allo stesso tempo, la deterrenza nucleare funziona solo fra attori con potenza militare simmetrica: uno Stato privo di armi nucleari non può esercitare deterrenza nei confronti di una potenza nucleare in questi termini.

A differenza della deterrenza nucleare, la cyberdeterrenza è ripetibile, perché è improbabile che le rappresaglie non cinetiche sconfiggano definitivamente l'avversario, e certo non costituiscono minacce definitive (Libicki, 2009). L'aggressore pertanto è in grado di compiere controrappresaglie e questo favorisce il moltiplicarsi delle interazioni tra chi difende e chi attacca. Le prime analisi (*ibidem*) sostengono che la cyberdeterrenza fra Stati è simmetrica, perché avviene tra pari ed entrambe le parti condividono lo stesso terreno strategico. Questo è solo parzialmente corretto, perché esistono scenari in cui chi si difende può avere capacità cibernetiche inferiori e per rappresaglia può usare mezzi cinetici (proporzionati), o in cui chi attacca si basa su mezzi cibernetici per colpire un avversario con mezzi cinetici superiori. Questo è il caso che descrive Geers: "Poiché la guerra cibernetica è guerra non convenzionale e asimmetrica, è probabile che le nazioni con scarsa potenza militare convenzionale vi investano per compensare gli svantaggi nella forza convenzionale" (2012, p. 5). L'uso non simmetrico delle capacità cibernetiche è stato riconosciuto anche in un report trapelato dalla NSA, in cui si ammette che "i cyberattacchi offrono, a potenziali avversari, il modo di compensare i vantaggi schiaccianti degli USA nella potenza militare convenzionale e di farlo in modi che sono istantanei e straordinariamente difficili da rintracciare" (National Security Agency, 2012, p. 3). Anche se ci si concentra solo sugli Stati, non è possibile presupporre una simmetria fra le capacità cibernetiche di un attaccante e di un difensore. Se ne conclude che la cyberdeterrenza è non simmetrica. Si tratta di un aspetto cruciale, perché significa che, nel decidere se rispondere o no con una rappresaglia, chi si difende dovrà prendere in considerazione la possibilità di una controrappresaglia sia cinetica sia non cinetica, e quindi di un'escalation. I due binomi – singolare e simmetrica o ripetibile e non simmetrica – indicano che la deterrenza nucleare e quella

cibernetica non sono collegate e, pertanto, le analogie non hanno un fondamento sicuro.

Le strategie di deterrenza inoltre sono fortemente determinate dalla natura delle minacce che pongono e di quelle che vogliono evitare. Nella deterrenza nucleare, la natura esistenziale delle minacce giustifica e rende credibili le strategie MAD. La cyberdeterrenza invece permette a chi si difende una gamma di strategie possibili – da una rappresaglia della stessa natura a sanzioni economiche, da misure diplomatiche a risposte cinetiche proporzionate – per la natura non esistenziale delle minacce cibernetiche non cinetiche. Queste opzioni vanno perse se si sceglie un modello della cyberdeterrenza in analogia con la deterrenza nucleare.

5.4.2.1 Controllo e rischi della cyberdeterrenza per rappresaglia

Anche se non si fa ricorso ad analogie con le strategie nucleari, la rappresaglia quale è identificata nella teoria della deterrenza pone seri problemi se applicata come strategia di deterrenza nel cyberspazio. A differenza della difesa, non è detto che la deterrenza per rappresaglia sia inefficace in questo ambito. In effetti, in un ambiente *offence-persistent* come il cyberspazio, la rappresaglia può essere una strategia di successo. Quando viene messa in atto nel cyberspazio, però, la natura delle armi cibernetiche e del conflitto cibernetico mina l'elemento di controllo della rappresaglia, rendendola una scelta pericolosa per la deterrenza.

La rappresaglia è accompagnata dal rischio di escalation, rischio che si amplifica quando la rappresaglia avviene in uno scenario non simmetrico, in cui l'avversario può non avere capacità cibernetiche adeguate e pertanto sceglierà una controrappresaglia con mezzi cinetici. Il controllo sui mezzi di rappresaglia e sul loro impatto è fondamentale per evitare un simile rischio. Nel cyberspazio, però, questo controllo è limitato, data la *malleabilità* delle armi cibernetiche, che fa sì che queste armi possano essere acquisite, combinate, convertite e riutilizzate molto più facilmente di quanto non sia mai stato possibile con altri tipi di mezzi militari (Schneier, 2017). Non è raro incontrare un malware progettato o fatto proprio da uno Stato, riconvertito e riutilizzato. È successo nel 2011 con Stuxnet, il famoso *worm* usato per attaccare gli impianti nucleari iraniani. Nonostante fosse stato progettato per mirare specificamente a determinati requisiti di configurazione del software Siemens installato nelle

centrifughe nucleari iraniane, da allora è stato riconvertito e utilizzato per attaccare sistemi in Azerbaigian, Indonesia, India, Pakistan e Stati Uniti.[4] Ancora più preoccupante è il fatto che la vulnerabilità sfruttata da Stuxnet era stata utilizzata per armare Angler, uno dei malware più insidiosi impiegati da criminali informatici per colpire siti web di online banking.[5] Analogamente, nel 2017 due forti cyberattacchi, WannaCry e NotPetya, hanno riconvertito un exploit (EternalBlue) trafugato dalla NSA.[6]

La probabilità che un'arma cibernetica provochi più danni di quelli pianificati in origine aumenta se si considera il probabile utilizzo, per la difesa nazionale, di *counter autonomy systems*, sistemi che mirano a neutralizzare o disabilitare le capacità decisionali e operative di sistemi autonomi e possono identificare e prendere a bersaglio le vulnerabilità di altri sistemi e al contempo isolare e rimediare alle proprie.[7] Dato che i sistemi IA apprendono e perfezionano il proprio comportamento attraverso le interazioni con l'ambiente (si ricordi il problema della predicibilità di cui abbiamo parlato nel capitolo 1), il loro uso a fini di difesa genera rischi concreti di danni imprevisti e sproporzionati.

La malleabilità delle armi cibernetiche, combinata con le capacità dell'IA, erode l'elemento di controllo della rappresaglia nel cyberspazio e, così facendo, rende la rappresaglia una scelta strategica pericolosa, con potenziali effetti a cascata disastrosi. Un controllo debole sull'impatto della rappresaglia può portare a una violazione della proporzionalità, che a sua volta può attivare risposte e controrappresaglie da parte dell'attaccante e favorire l'escalation. È essenziale garantire il controllo sulla rappresaglia per evitare questi effetti non voluti, e a tal fine è essenziale rispettare la proporzionalità. Come scrive Iasiello:

> Uno Stato-nazione non deve solo rispondere all'aggressore, ma deve farlo in modo da chiarire la propria posizione – deve essere, cioè, una risposta potente, ma non tanto potente da suscitare una reazione negativa nella comunità globale. (2014, p. 59)

Esiste un consenso generale sul fatto che il principio di proporzionalità si applichi nel caso della cyberdeterrenza e la rappresaglia per un cyberattacco può essere affidata a mezzi cibernetici o cinetici (o a una combinazione dei due), purché la risposta sia paragonabile per impatto all'attacco iniziale e non provochi un'escalation (Libicki, 2009; Jensen, 2009; Goodman, 2010; Hathaway, Crootof, 2012; Iasiello, 2014).

Come abbiamo visto nel capitolo 4, però, determinare l'impatto di un cyberattacco non cinetico è problematico. Il principio di proporzionalità afferma che la rappresaglia deve essere corrispondente al danno reale subito (e non solo a quello scoperto). Questo può essere un ostacolo grave nel cyberspazio, dove "molto poco […] si può inferire a proposito di attività non viste (che non si possono misurare) da quelle che si vedono (che possono essere misurate)" (Libicki, 2009, p. 103). Al contempo, anche se l'attacco è stato individuato e il suo impatto è chiaro, può risultare difficile stabilire il valore e il tipo di danno, perciò anche l'entità di una risposta appropriata. Come nota Harknett:

> Se un attacco non riduce in macerie alcun edificio e non uccide direttamente nessuno, ma distrugge informazione, qual è la risposta? Tendiamo a pensare che l'informazione sia qualcosa di intangibile, ma la perdita di informazione può avere costi tangibili a livello personale, istituzionale e sociale. Che cosa si potrebbe mettere credibilmente a rischio, che possa dissuadere uno Stato dal prendere in considerazione un attacco del genere? (1996, p. 104)

Le risposte a queste domande dipendono dalla lacuna ontologica fra Teoria della Guerra Giusta e guerra cibernetica discussa nel capitolo 4 e richiedono un rimodellamento della teoria della deterrenza, che tenga conto della natura del cyberspazio e della guerra cibernetica non cinetica.

In scenari cinetici, le strategie di difesa e rappresaglia offrono l'equilibrio perfetto tra controllo della risposta e coercizione che in ultima istanza consentono a chi si difende di mostrare la propria forza e di scoraggiare l'attaccante. Nel cyberspazio, un equilibrio del genere non si può ottenere, perché né la difesa né la rappresaglia sono controllabili, ed entrambi i tipi di deterrenza diventano impraticabili. Tuttavia, bisogna notare una differenza importante: mentre è certo che le strategie difensive non sono efficaci in un ambiente *offence-persistent* come il cyberspazio, la rappresaglia può essere un'opzione strategica percorribile (nei limiti imposti da attribuzione e proporzionalità).

Ciononostante, per avere successo, la rappresaglia va riconsiderata per garantire che, pur rimanendo essenzialmente un atto di coercizione, possa basarsi su meccanismi robusti di controllo, che garantiscano una risposta proporzionata. Per questo è necessario un quadro normativo, che sarà efficace solo nella misura in cui colmi la lacuna ontologica e superi i limiti della Teoria della Guerra Giusta nel cyberspazio (tornerò su questo punto nel paragrafo 5.6). L'alternativa è modellare la rappresaglia nel

cyberspazio con strategie MAD, che però porterebbero all'escalation, anziché a prevenire nuovi conflitti.

5.5 SEGNALAZIONE CREDIBILE

Un difensore cerca di dissuadere i possibili attaccanti segnalando di essere consapevole dei loro piani e di avere pronta una risposta, nel caso quei piani venissero attuati. Senza questi segnali, la deterrenza non sarebbe possibile. Iasiello, per esempio, nota che la rappresaglia diventa inefficace e può essere interpretata in modo errato se chi si difende non è in grado di inviare un segnale credibile delle proprie intenzioni (Iasiello, 2014). Come abbiamo visto nel modello D_M, la segnalazione è efficace solo nella misura in cui trasmette un messaggio coercitivo (minaccia) e, pertanto, dipende dalla messa in atto di una strategia di deterrenza appropriata (vedi Figura 5.1). Il messaggio deve essere credibile, e la sua credibilità dipende dal fatto che il difensore goda della reputazione di chi dà seguito alle proprie minacce (Freedman, 2004). In effetti, la reputazione è un aspetto centrale della teoria della deterrenza. Schelling ha sottolineato (in un passo famoso) che "la faccia è una delle poche cose per cui vale la pena di combattere […] 'la faccia' è semplicemente l'interdipendenza degli impegni di un paese; è la reputazione di un paese per le sue azioni, le aspettative che altri paesi hanno in merito al suo comportamento" (*ibidem*, p. 53).

In scenari cinetici, la reputazione si guadagna esibendo le capacità militari dello Stato – parate militari e schieramento di soldati o navi ai confini dello Stato offensore in genere servono a questo scopo – e anche mostrando la capacità di dissuadere e di sconfiggere gli avversari nel tempo. In una certa misura, vale lo stesso anche per il cyberspazio, dove la reputazione di uno Stato dipende dalle interazioni passate in quest'ambiente, dalle sue capacità cibernetiche note per la difesa e l'offesa, così come dalla sua reputazione generale nella risoluzione di conflitti. Bisogna stare attenti, però, perché la reputazione di uno Stato nel cyberspazio può non corrispondere alle sue capacità effettive, visto che per esempio gli Stati sono riluttanti a divulgare informazioni sugli attacchi che ricevono. Nel medio e lungo termine, questo può rendere la segnalazione meno credibile, e quindi più problematica, che in altri ambiti della difesa.

La segnalazione può essere o generica o “su misura”. La segnalazione generica trasmette un messaggio sulla strategia generale di deterrenza al resto dell’arena internazionale, attraverso dichiarazioni pubbliche da parte di uno Stato che veicolano informazioni sui suoi approcci, i suoi impegni e le sue capacità. Anche se può essere difficile in alcune circostanze, la segnalazione generica nel cyberspazio non è impossibile. Per esempio, i riferimenti alla capacità di ricorrere alla “gamma completa di strumenti disponibili agli Stati Uniti” nel documento di strategia cibernetica degli USA (US Government, 2015, p. 14), come la menzione delle capacità di Active Cyber Defence nell’equivalente del Regno Unito (UK Government, 2015), servono a questo scopo. In entrambi i casi, la segnalazione generica è credibile e si basa sulla reputazione di cui godono nel cyberspazio Stati Uniti e Regno Unito.

Il *tailored signalling* – cioè la comunicazione di una minaccia mirata a un attore specifico, con l’indicazione dei possibili obiettivi di una ritorsione – presenta criticità ben superiori rispetto alla segnalazione generica e costituisce un ostacolo rilevante all’efficacia delle strategie di deterrenza nel cyberspazio. Questo tipo di segnalazione funziona solo se l’attribuzione è certa: in assenza di una chiara identificazione dell’aggressore, un messaggio mirato rischia di colpire il bersaglio sbagliato, risultando inefficace o addirittura controproducente. Inoltre, la segnalazione deve essere calibrata con attenzione, per evitare che il difensore riveli capacità e asset sensibili, specialmente qualora stia valutando una ritorsione simmetrica.

I rischi sono molteplici e vanno dall’esposizione del proprio grado di conoscenza delle capacità cibernetiche dell’avversario – con la conseguente implicazione che siano già state condotte operazioni offensive, di sabotaggio o spionaggio – fino alla compromissione di strumenti e strategie del difensore stesso, come vulnerabilità *zero-day*, che una volta note perdono il loro valore operativo. Al tempo stesso, un messaggio troppo vago minerebbe la credibilità della minaccia e quindi l’efficacia della deterrenza.

Da ciò derivano due possibili sviluppi per la deterrenza nel cyberspazio: o la segnalazione si dissocia progressivamente dalla reputazione dell’attore, indebolendo così la dimensione coercitiva della deterrenza stessa; oppure la deterrenza si allontana dall’idea di *signalling* come avvertimento preventivo e si avvicina a una logica di dimostrazione

concreta di capacità e intenzioni. Ma quest’ultima strategia comporta un rischio crescente di escalation. Entrambi gli scenari compromettono le premesse fondamentali per l’efficacia delle strategie deterrenti basate sulla teoria classica della deterrenza.

5.6 IA PER LA CYBERDETERRENZA: UN NUOVO MODELLO

Dall'analisi dei limiti della teoria della deterrenza segue che, perché la deterrenza funzioni nel cyberspazio, la minaccia condizionale e gli elementi coercitivi, che nella teoria sono associati, devono essere separati. Questo perché l'elemento minaccia è vanificato dai limiti dell'attribuzione, dalla natura *offence-persistent* del cyberspazio e dai rischi legati alla credibilità della segnalazione.

Tenendo conto di tutto questo, presento un nuovo modello per la cyberdeterrenza, che si basa su tre elementi centrali: identificazione del bersaglio, rappresaglia e dimostrazione (Figura 5.3). Approfondiamoli.

Figura 5.3 I tre elementi della teoria della cyberdeterrenza e le loro dipendenze (Taddeo, 2018a, p. 5).

Secondo questo modello, l'identificazione del bersaglio e la rappresaglia hanno uno scopo dimostrativo, perché nel cyberspazio l'attacco è la scelta razionale, dato che per l'avversario le probabilità di successo sono elevate e quelle di una punizione basse. In questo scenario, le *minacce* di rappresaglia non sono un deterrente per un avversario che già sta preparando un attacco, che ha acquisito intelligence e capacità, e ha

approntato un piano. Allo stesso tempo, nella misura in cui può esporre le capacità cibernetiche, minacciare un contrattacco non è una strategia praticabile per chi si difende.

Per essere efficace, la rappresaglia deve mettere in mostra le capacità e le intenzioni del difensore e causare all'avversario danni sufficienti a vanificare l'esito di un attacco andato a buon fine. La deterrenza, di conseguenza, non mira a evitare gli attacchi in arrivo, ma a modificare i calcoli che potrebbero indurre lo stesso avversario ad attaccare di nuovo in futuro. La cyberdeterrenza punta a evitare la *successiva* ondata di attacchi condotti dallo stesso avversario. Per avere successo nel cyberspazio, la deterrenza deve passare dal minacciare al prevalere. Dimostrare competenze e capacità per identificare e attaccare le risorse dell'avversario è il modo più efficace di prevalere. A questo scopo, l'identificazione del bersaglio è cruciale.

In questo modello, l'identificazione del bersaglio sostituisce l'attribuzione. Come abbiamo visto nel paragrafo 5.3, nel cyberspazio un'attribuzione positiva può essere problematica, anche se è vero che, nel corso degli anni, è diventata sempre più possibile. A mano a mano che gli Stati acquisiscono maggiori informazioni sui loro avversari, e la posizione e le tattiche di alcuni attori nel cyberspazio diventano più chiare, attribuire correttamente gli attacchi è diventato più facile. L'IA è di grande rilevanza a questo scopo, perché è possibile utilizzarla per analizzare grandi quantità di dati (vedi il capitolo 3) e raccogliere intelligence a sostegno dell'identificazione degli attaccanti, per esempio attraverso la loro profilazione (Chen, 2016). Noor e colleghi propongono di utilizzare l'elaborazione del linguaggio naturale e l'apprendimento profondo per

> profilare attori di minacce cibernetiche (*cyber threat actors*, CTA) sulla base dei loro schemi d'attacco estratti da report di intelligence sulle minacce, utilizzando [l'elaborazione del linguaggio naturale]. Grazie a questi profili, addestriamo e sottoponiamo a test cinque classificatori di *machine learning* su 327 report di intelligence su minacce cibernetiche elaborati in base alla documentazione disponibile pubblicamente su incidenti avvenuti fra il maggio 2012 e il febbraio 2018. Si osserva che i profili degli attori così ottenuti attribuiscono le minacce cibernetiche con una precisione elevata (83%, mentre altri profili CTA disponibili pubblicamente hanno una precisione del 33%). Il classificatore basato sulla Deep Learning Neural Network inoltre attribuisce le minacce cibernetiche con un'accuratezza ancora superiore (94%, rispetto ad altri classificatori). (2019, p. 227)

Vale la pena di sottolineare che l'identificazione degli attaccanti non equivale all'attribuzione. Nel contesto della guerra cibernetica, la seconda

implica un collegamento causale fra uno Stato attore e un cyberattacco per giustificare risposte geopolitiche e militari nei confronti di quello Stato. Nel cyberspazio, l'identificazione degli attaccanti può non essere sufficiente per attribuire un attacco a uno Stato attore, dato che il legame fra attaccanti e Stato può essere difficile da dimostrare. È il motivo per cui in questo modello distinguo fra identificazione del bersaglio e attribuzione, e mi concentro sulla prima. Identificazione del bersaglio significa identificare le risorse degli attaccanti e farne il bersaglio di una rappresaglia, senza dover dimostrare il collegamento fra gli attaccanti e gli Stati che li sostengono. L'identificazione del bersaglio è più specifica dell'attribuzione, perché comprende una valutazione di proporzionalità appropriata per evitare l'escalation, e una valutazione della robustezza delle risorse e delle misure difensive per garantire che la rappresaglia sia efficace. In questo senso, concentrarsi sull'identificazione del bersaglio anziché sull'attribuzione facilita la deterrenza.

Questo modello della cyberdeterrenza non prevede la difesa tra le sue strategie possibili, come si vede nella Figura 5.3. Ciò dipende dalla natura *offence-persistent* del cyberspazio, e non implica che non si debba prendere in considerazione la difesa nel cyberspazio, ma semplicemente che in questo ambiente non funziona come deterrente. Vale la pena di sottolineare che la deterrenza basata sulla rappresaglia può produrre rischi gravi per la stabilità del cyberspazio. Il controllo dei suoi effetti in questo ambiente è problematico e l'uso dell'IA per metterla in atto rende il controllo ancora più difficile. Questi problemi, però, possono essere affrontati con successo con iniziative normative per regolare il comportamento degli Stati nel cyberspazio, compreso l'uso dell'IA per finalità conflittuali e non cinetiche.

Come abbiamo visto in precedenza (vedi i capitoli 1 e 4), l'uso dell'IA per scopi conflittuali e non cinetici introduce rischi gravi di escalation delle risposte, conseguenti dalle capacità di apprendimento di questa tecnologia e dalla sua predicibilità limitata. Per sfruttare il potenziale dell'IA per la deterrenza e la stabilità nel cyberspazio è fondamentale che il suo uso rispetti i principi etici presentati nei capitoli 2 e 4. Questo comporta la necessità di specificare regole per il comportamento degli Stati nel cyberspazio e per il ciclo di vita delle tecnologie IA utilizzate per scopi conflittuali e non cinetici.

Per quanto riguarda la regolamentazione del comportamento degli Stati, vale la pena di sottolineare che il modello della cyberdeterrenza presentato qui rimane coerente con i principi (P1-P3) per la guerra cibernetica discussi nel capitolo 4, purché la rappresaglia abbia effetti non cinetici. P1 prevede l'eliminazione di qualsiasi entità che possa causare entropia nell'infosfera; se la rappresaglia prende a bersaglio sistemi utilizzati (o in procinto di essere utilizzati) per condurre un cyberattacco, è giustificata in base a questo principio. P1-P3 richiedono che le risposte siano proporzionate al male, perciò nel condurre una rappresaglia non bisogna generare maggiore entropia metafisica (danno) di quella che si intende eliminare. Questi principi integrano quelli della Teoria della Guerra Giusta, che rimangono validi quando si considerano operazioni conflittuali e non cinetiche condotte o sponsorizzate da Stati. La rappresaglia nel cyberspazio comporta quindi una valutazione appropriata di proporzionalità, necessità e distinzione. Finché le risposte rimangono non cinetiche, attacchi nel cyberspazio restano preferibili a quelli cinetici, ma questo è vero solo se si definiscono con chiarezza vincoli e regole per evitare l'escalation e conservare il controllo degli effetti.

Per questo è importante che venga stabilito un regime internazionale di norme che regolano il comportamento degli Stati nel cyberspazio, a complemento delle strategie nazionali di cyberdeterrenza e per favorire la stabilità. A mano a mano che l'IA diventa una capacità centrale per la cyberdeterrenza, quelle norme dovranno tenere conto delle caratteristiche tecniche dei sistemi IA e prevedere e mettere in atto misure appropriate per mitigare i rischi relativi. A questo scopo sono fondamentali quattro passi.

- Definire linee chiare di distinzione fra bersagli legittimi e non, e definire risposte proporzionate per la cyberdeterrenza e le strategie di difesa.
- Costituire alleanze imponendo esercizi di conflitto amichevole fra alleati per testare le capacità basate su IA e scoprire le vulnerabilità fatali di sistemi chiave e infrastrutture indispensabili fra alleati.
- Definire standard internazionali per valutare la predicibilità dei sistemi IA usati nella difesa, e definire soglie di tolleranza.
- Monitorare e far valere regole a livello internazionale definendo procedure di auditing e supervisione delle operazioni di difesa

> cibernetica degli Stati basate su IA, prevedendo meccanismi di allerta e riparazione per rimediare a errori e conseguenze non volute.

Una volta definito e concordato, questo regime dovrà essere rispettato. Ciò richiederà un'autorità indipendente, in grado di esercitare un potere coercitivo e di imporre punizioni. Questa autorità non può (e non dovrebbe) essere il risultato di un'iniziativa multistakeholder o di un'iniziativa neutrale, guidata da privati.[8] Ciò imporrebbe responsabilità troppo grandi al settore privato, e al tempo stesso creerebbe un'autorità troppo debole per resistere alla pressione politica derivante dal dover garantire il rispetto di un regime cibernetico internazionale da parte degli Stati.

Per il buon funzionamento di questo regime è necessaria un'autorità in grado di (i) verificare che gli Stati rispettino le norme, (ii) intraprendere indagini su sospetti cyberattacchi condotti (o sponsorizzati) da Stati per accertare l'identificazione delle fonti di quegli attacchi e, se possibile, anche la loro attribuzione, (iii) esporre le violazioni delle norme e le fonti di cyberattacchi illegittimi, e (iv) imporre sanzioni o punizioni adeguate. Per raggiungere questi obiettivi sono necessari il coordinamento di capacità di intelligence, politiche e diplomatiche, e competenze tecniche estremamente avanzate, nonché l'autorità politica e un apparato per imporre il rispetto delle sanzioni. Le capacità (i)-(iv) definiscono un mandato politico per un'autorità che avrà un impatto profondo sulle relazioni internazionali e gli equilibri geopolitici.

Questo è in perfetto accordo con l'Articolo 26 dello Statuto delle Nazioni Unite, che definisce il mandato del Consiglio di Sicurezza:

> Al fine di promuovere lo stabilimento ed il mantenimento della pace e della sicurezza internazionale col minimo dispendio delle risorse umane ed economiche mondiali per gli armamenti, il Consiglio di Sicurezza ha il compito di formulare, con l'ausilio del Comitato di Stato Maggiore […] piani da sottoporre ai Membri delle Nazioni Unite per l'istituzione di un sistema di disciplina degli armamenti.[9]

Il Consiglio di Sicurezza delle Nazioni Unite deve quindi assumersi il compito di istituire e sostenere una tale autorità. In un circolo vizioso, cyberattacchi, usi conflittuali e non cinetici dell'IA e la corsa agli armamenti cibernetici si alimentano a vicenda – ponendo una grave

minaccia alla stabilità del cyberspazio e quindi alla sicurezza e alla pace delle nostre società digitali. Le strategie offensive da sole hanno fallito, e continueranno a fallire, nel tentativo di spezzare questo circolo, ma possono avere successo se affiancate alle appropriate misure normative. Il Consiglio di Sicurezza delle Nazioni Unite ha le risorse necessarie, compreso il potere politico e coercitivo, per guidare e implementare questo processo.

5.7 CONCLUSIONE

La teoria della deterrenza incontra seri limiti quando è applicata al cyberspazio. Sarebbe un errore, però, concludere da tali limiti che la deterrenza non è possibile in questo ambiente. Come ha affermato il comandante della marina USA Bebber (2018):

> La storia ci dice che applicare il quadro operativo sbagliato a un ambiente strategico emergente è una ricetta sicura per il fallimento. Durante la Prima guerra mondiale, entrambe le parti non si resero conto che i bombardamenti di artiglieria su larga scala seguiti da assalti di fanteria in massa erano inutili su un campo di battaglia che favoriva fortemente una difesa ben trincerata, sostenuta dalla tecnologia delle mitragliatrici. […] L'incapacità di adattarsi ebbe conseguenze disastrose.

Dobbiamo adattare il nostro modo di pensare sulla base di una comprensione approfondita della guerra non cinetica, della sua natura e della sua dinamica, per creare un nuovo modello di deterrenza in grado di affrontarla. L'alternativa – sviluppare la cyberdeterrenza per analogia con la deterrenza convenzionale – è una ricetta per il fallimento.

Nel 2017, i ministri degli Esteri dei paesi del G7 (cioè Canada, Francia, Germania, Italia, Giappone, Regno Unito e Stati Uniti) hanno approvato una Dichiarazione sul comportamento responsabile degli Stati nel cyberspazio (G7 Declaration, 2017). La Dichiarazione affronta la preoccupazione crescente per la stabilità internazionale e la sicurezza delle nostre società dopo la rapida escalation dei cyberattacchi non cinetici negli ultimi anni. Nelle frasi di apertura, i ministri del G7 sottolineano la preoccupazione per

> il rischio di escalation e rappresaglia nel cyberspazio. […] Tali attività possono avere un effetto destabilizzante sulla pace e la sicurezza a livello internazionale. Sottolineiamo che il rischio di conflitti fra Stati in conseguenza di incidenti ICT è emerso come un problema che è urgente prendere in considerazione. (*Ibidem*, p. 1)

Negli ultimi anni, vari attori nazionali, internazionali e sovranazionali – la NATO (Freedberg, 2014), l'Institute for Disarmament Research delle Nazioni Unite (UN Institute for Disarmament Research, 2014), il governo del Regno Unito (UK Government, 2014) e il Dipartimento di Stato degli Stati Uniti (International Security Advisory Board, 2014) – hanno ribadito

la necessità di definire strategie per garantire la stabilità di fronte a una tendenza all'escalation dei cyberattacchi. Paradossalmente, gli attori statali hanno spesso un ruolo centrale in questa escalation. Sono stati lanciati cyberattacchi condotti e sponsorizzati da Stati a fini di spionaggio e sabotaggio almeno dal 2003. Titan Rain (2003), l'attacco russo contro l'Estonia (2006) e la Georgia (2008), Red October (2007), Stuxnet e Operation Olympic Game contro l'Iran (2006-2012) sono esempi ben noti. Casi famosi sono anche quelli del cyberattacco russo contro una centrale energetica ucraina,[10] le infiltrazioni cinesi e russe degli uffici federali degli Stati Uniti,[11] i cyberattacchi Shamoon/Greenbag alle infrastrutture del governo in Arabia Saudita,[12] e la campagna di cyberattacchi condotti o sponsorizzati a livello statale che ha preso di mira infrastrutture e servizi digitali europei dopo l'invasione russa dell'Ucraina.[13]

Questa tendenza continuerà, con rischi sempre più gravi di escalation. Lo sviluppo di strategie di deterrenza che possano affrontare la natura della guerra cibernetica è fondamentale, ma non è sufficiente a mitigare questi rischi. Qui l'IA ha un ruolo centrale. È fondamentale che queste strategie siano accompagnate da regolamentazioni del comportamento degli Stati, fatte rispettare da un'autorità dotata di poteri effettivi. Si tratta di uno sforzo complesso, ma anche necessario, in considerazione del livello a cui le società digitali dipendono dalle loro risorse digitali. Questo capitolo, insieme con il precedente, ha avuto l'obiettivo di mostrare come analisi concettuali ed etiche possano sostenere l'impegno a definire strategie di deterrenza efficaci che, unite a una regolamentazione appropriata, migliorino la stabilità del cyberspazio e delle società digitali che dal cyberspazio dipendono per il loro buon funzionamento.

---

1. L'attribuzione può non essere necessaria in tutti i casi di deterrenza, per esempio per la deterrenza mediante difesa. Qualcuno sostiene che, quando l'origine precisa di un attacco non è nota, l'attribuzione e, di conseguenza, la responsabilità di un attacco può essere fatta ricadere sul particolare Stato in cui l'attacco ha avuto origine (Morgan, 2010; Goodman, 2010). Tuttavia, un'attribuzione chiara rimane necessaria per la deterrenza mediante rappresaglia.
2. https://www.theregister.co.uk/2017/06/28/petya_notpetya_ransomware/.
3. http://www.telegraph.co.uk/technology/2017/05/23/highly-likely-wannacry-cyber-attack-linked-north-korea/.
4. https://www.symantec.com/security_response/writeup.jsp?docid=2010-071400-3123-99.

5. https://www.theregister.co.uk/2016/05/09/sixyearold_patched_stuxnet_hole_still_the_webs_biggest_killer/.

6. https://www.forbes.com/sites/thomasbrewster/2017/05/12/nsa-exploit-used-by-wannacry-ransomware-in-global-explosion/#3f04a279e599.

7. https://fas.org/irp/agency/dod/dsb/autonomy-ss.pdf; https://www.darpa.mil/program/cyber-grand-challenge.

8. Vedi per esempio la proposta di una "Convenzione di Ginevra digitale" avanzata da Microsoft nel 2017 per garantire che i governi proteggano i civili da cyberattacchi condotti o sponsorizzati da Stati: https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/#d4uIGGAJo1rg7Thg.99.

9. https://www.un.org/en/about-us/un-charter/chapter-5 (in italiano: https://it.wikisource.org/wiki/Statuto_delle_Nazioni_Unite).

10. https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/.

11. https://www.nytimes.com/2016/12/13/us/politics/russia-hack-election-dnc.html?_r=0.

12. https://www.symantec.com/connect/blogs/greenbug-cyberespionage-group-targeting-middle-east-possible-links-shamoon.

13. https://www.ncsc.gov.uk/news/russia-behind-cyber-attack-with-europe-wide-impact-hour-before-ukraine-invasion.

6

USI CONFLITTUALI E CINETICI DELL'IA

LA DEFINIZIONE DI SISTEMI D'ARMA AUTONOMI

6.1 INTRODUZIONE

Parliamo ora dell'uso dell'IA per scopi conflittuali e cinetici. Qui la discussione è centrata sui sistemi d'arma autonomi (*autonomous weapons systems*, AWS), sull'ammissibilità morale e sulla regolamentazione del loro uso. Il dibattito sulle implicazioni etiche e giuridiche degli AWS risale ai primi anni Duemila, quando alcuni (Arkin, 2009) difendevano l'uso di questi sistemi e altri invece ne volevano una completa messa al bando (Sharkey, 2008, 2010; Sparrow, 2007). Il dibattito è entrato nel vivo nel 2012, quando il DoD degli Stati Uniti ha pubblicato un ordine esecutivo sugli AWS (US Department of Defense, 2012) e Human Rights Watch ha condiviso un report sui problemi etici e giuridici sollevati dagli AWS ("Losing humanity", 2012). Da allora, la discussione si è sviluppata con contributi di studiosi, esperti militari e decisori politici e con il coinvolgimento dell'International Committee of the Red Cross (ICRC), dello UN Institute for Disarmament Research (UNIDIR) e della UN Convention on Certain Conventional Weapons (CCW). Quest'ultima ha istituito un Governmental Group of Experts (GGE) per definire delle regolamentazioni per le tecnologie emergenti nell'ambito dei sistemi d'arma autonomi letali (*lethal autonomous weapon systems*, LAWS).

Fino a oggi il GGE non è riuscito a definire le norme per cui è stato creato. Torneremo su questo punto nel capitolo 8, per ora è importante notare che il dibattito sia sulle implicazioni etiche sia su quelle legali dell'uso degli AWS è profondamente polarizzato. Esiste, però, almeno un consenso sugli aspetti etici e giuridici che devono essere presi in considerazione quando si tenta di valutare se l'uso di queste armi sia eticamente e legalmente accettabile: rispetto per la dignità umana, per i principi della Teoria della Guerra Giusta e dell'IHL, e impatto sulla stabilità internazionale. L'IHL è centrale in questo dibattito, perché tutti concordano sul fatto che gli AWS possono essere utilizzati solo se il loro uso ne rispetta i principi di necessità, proporzionalità e distinzione. Questi principi non sono in discussione; l'aspetto problematico è capire se, e in quale misura, gli agenti artificiali autonomi che caratterizzano gli AWS possano essere usati in modo conforme a quei principi.[1] Per esempio, rispettare il principio della distinzione per gli AWS è problematico perché

(almeno allo stadio attuale di sviluppo) gli agenti artificiali autonomi non sono in grado di analizzare il contesto in cui operano con la precisione necessaria per distinguere che cosa o chi sia un bersaglio legittimo (Sharkey, 2010, 2016; Amoroso, Tamburrini, 2020). In tal senso alcuni sostengono che i principi dell'IHL definiscono requisiti operativi che, se non sono soddisfatti dai modelli attuali degli AWS, almeno in linea di principio potrebbero essere soddisfatti in futuro da sistemi più avanzati. Nel capitolo 8 vedremo che questo approccio non tiene conto di alcune limitazioni intrinseche alle tecnologie dell'IA.

Problemi più fondamentali emergono quando si considerano gli AWS e la dignità umana. In questo caso l'interrogativo riguarda *come* una persona viene uccisa o ferita. In altre parole, l'attenzione qui è rivolta al processo con il quale vengono prese le decisioni di uccidere o colpire un essere umano. Se la decisione è presa da una macchina, allora si prospetta la violazione della dignità umana dei bersagli (Asaro, 2012; Docherty, 2014; Sharkey, 2019; Johnson, Axinn, 2013; Sparrow, 2016; O'Connell, 2014; Ekelhof, 2019). L'impatto dell'uso degli AWS sulla dignità umana è indipendente dal livello di sofisticazione della tecnologia, perché il problema in questo caso riguarda la legittimità di delegare la decisione sull'uso della forza (letale) a macchine (Lieblich, Benvenisti, 2016) e se delegare questa decisione sia compatibile con i valori sostenuti dalle nostre società. I temi di umanità e coscienza pubblica, che sono centrali per valutare la legittimità di qualsiasi arma, sono cruciali anche quando si considerano gli AWS. Come sottolinea il report dell'ICRC:

> Le decisioni etiche prese dagli Stati, e dalla società in generale, hanno preceduto e motivato lo sviluppo di nuovi vincoli giuridici nella guerra, compresi vincoli sulle armi che causano danni non accettabili. Nel diritto umanitario internazionale, le nozioni di umanità e coscienza pubblica sono derivate dalla clausola Martens. (International Committee of the Red Cross, 2018, p. 1)

In ultima analisi, le questioni relative alla dignità umana rimandano alla *human agency* (vale a dire la capacità di scegliere, decidere e agire) e alle decisioni e azioni che si possono – o non si possono – delegare alle macchine, e alle responsabilità morali che ne derivano, in particolare quando è in gioco l'uso della forza. Attribuire la responsabilità morale per le scelte compiute dai sistemi IA si è rivelato estremamente problematico in numerosi ambiti, e il caso degli AWS non fa eccezione. Come ho sostenuto altrove (Taddeo et al., 2021), sebbene il *responsibility gap*

rappresenti una criticità in tutte le categorie di impiego dell'IA nel settore della difesa – dal supporto logistico alle operazioni conflittuali non cinetiche, fino a quelle cinetiche, esso appare particolarmente allarmante negli usi cinetici, dove la posta in gioco è massima (Sparrow, 2007).

L'uso degli AWS potrebbe avere anche un impatto negativo sulla stabilità internazionale. Qualcuno ha sostenuto che possono abbassare le barriere al conflitto armato, portando a una maggiore incidenza della guerra e ostacolando la stabilità internazionale (Enemark, 2011; Brunstetter, Braun, 2013). Per esempio, un uso diffuso degli AWS potrebbe consentire ai decisori di intraprendere una guerra senza bisogno di rispondere alle obiezioni del personale militare o più in generale della popolazione in un contesto democratico (Steinhoff, 2013; Heyns, 2014). Analogamente, una guerra asimmetrica risultante dall'uso di AWS da parte di un aggressore potrebbe portare la parte più debole a ricorrere più spesso all'insurrezione e a tattiche terroristiche (Sharkey, 2012a, 2012b). Dato che il terrorismo in generale è considerato una forma di guerra ingiusta (o, peggio, un atto di omicidio indiscriminato), l'uso degli AWS può portare a una maggiore incidenza della violenza ingiusta.

L'attenzione per questi temi è cresciuta nel tempo, sia nell'ambito accademico sia in quello politico. Oltre un decennio dopo l'ordine esecutivo del DoD e il report di Human Rights Watch citati prima, però, non è ancora stato definito un approccio internazionale condiviso a questi problemi. Le ragioni dell'insuccesso sono molteplici e spaziano dalla mancanza di volontà politica ai conflitti d'interesse sul piano internazionale, fino alle posture difensive degli Stati. Il tutto è ulteriormente aggravato dall'assenza di una comprensione condivisa degli AWS, delle loro caratteristiche fondamentali e delle implicazioni etiche e giuridiche che ne derivano. Come è stato sottolineato in un report dell'UNIDIR, "tanto chi sostiene gli AWS quanto chi vi si oppone cercherà di formulare una definizione funzionale ai suoi obiettivi e ai suoi interessi. La discussione sulla definizione non sarà un esame di fatti, neutro rispetto ai valori, ma alla fine sarà guidata da motivazioni politiche e strategiche" (2017, p. 22).

In questo capitolo mi concentro sulle definizioni di AWS proposte da Stati o attori internazionali fondamentali come ICRC e NATO per offrirne un'analisi comparativa. Le dodici definizioni raccolte si focalizzano su aspetti diversi degli AWS e pertanto conducono ad approcci differenti

nell'affrontare i problemi etici e giuridici posti da questi sistemi d'arma. Ne deriva un approccio a dir poco frastagliato, spesso incoerente, non solo alla regolamentazione degli AWS, ma alla loro stessa definizione. La mancanza di coerenza nella definizione degli AWS influisce negativamente sia sulla possibilità di arrivare a una comprensione comune di questi sistemi, sia sul raggiungimento di un accordo sulla regolamentazione del loro uso e, di fatto, sul loro stesso uso in generale. Ciò diventa evidente quando si esamina il lavoro di CCW/GGE. La Tabella 6.1 riassume i punti chiave della discussione di questo gruppo fra 2014 e 2019. Si può vedere che, mentre esiste un consenso sugli aspetti fondamentali degli AWS e sui problemi etici che pongono, continua a mancare una definizione condivisa, e pertanto una comprensione condivisa degli AWS e di quali aspetti presentino i problemi etici e giuridici più urgenti. Per esempio, spesso le discussioni del CCW/GGE riportate nella Tabella 6.1 non distinguono tra AWS e LAWS e quindi non permettono di dare priorità ai relativi problemi etici e normativi.

Tabella 6.1 Punti chiave delle discussioni del CCW/GGE fra 2014 e 2019.

CCW/GGE	
2014	"Molti interventi hanno sottolineato il fatto che, anche se l'elaborazione di una definizione era prematura, alcuni elementi chiave apparivano pertinenti per la descrizione del concetto di autonomia per i LAWS, per esempio la capacità di selezionare e ingaggiare un bersaglio senza intervento umano. Alcuni esperti hanno evidenziato il fatto che l'autonomia deve essere misurabile e basata su criteri oggettivi, come la capacità di percezione dell'ambiente e la possibilità di eseguire compiti preprogrammati senza ulteriore azione umana. Molti interventi hanno sottolineato che la nozione di controllo umano significativo può essere utile per affrontare la questione dell'autonomia. Altre delegazioni hanno affermato anche che questo concetto richiede ulteriore studio nel contesto della CCW. È stato discusso anche il concetto di coinvolgimento umano in progettazione, test, revisioni, addestramento e uso. Alcune delegazioni hanno sottolineato che anche la nozione di predicibilità è un aspetto fondamentale" (Simon-Michel, 2014, p. 4).
2017	"È stata riconosciuta la necessità di migliorare la comprensione condivisa dei sistemi d'arma autonomi. È stata auspicata l'elaborazione di una definizione operativa di LAWS, senza pregiudizio per la definizione di sistemi che possano essere oggetto di regolamentazione futura. È stato preso in considerazione l'ambito di una possibile definizione, incluse questioni relative a sistemi già in uso, ad armi difensive o offensive e alla distinzione fra sistemi pienamente autonomi e semi-autonomi. È stata avanzata anche l'idea che fosse prematuro o poco utile iniziare a lavorare sulle definizioni" (Korpela, 2017, p. 7).

CCW/GGE	
2018	“Sono state ulteriormente studiate le caratteristiche tecniche relative all’autoapprendimento (senza dati di addestramento forniti dall’esterno) e all’autoevoluzione (senza input di progettazione umana). Analogamente, tentare di definire un livello soglia generale di autonomia in base a criteri esclusivamente tecnici potrebbe creare difficoltà, dato che l’autonomia è uno spettro, la sua interpretazione cambia con gli spostamenti della frontiera tecnologica e funzioni diverse di un sistema d’arma potrebbero avere gradi di autonomia diversi. […] Nel contesto della CCW, è necessario concentrarsi sulle caratteristiche relative all’elemento umano nell’uso della forza e alla sua interfaccia con le macchine, per affrontare i temi di *accountability* e responsabilità” (Convention on Certain Conventional Weapons, 2018, p. 5).
2019	“Sul punto 5(b) dell’agenda, ‘Caratterizzazione dei sistemi in considerazione, al fine di promuovere una comprensione comune di concetti e caratteristiche rilevanti agli obiettivi e agli scopi della Convenzione’, il Gruppo ha concluso quanto segue: (a) il ruolo e le conseguenze delle funzioni autonome in identificazione, selezione e ingaggio di un bersaglio sono fra le caratteristiche essenziali dei sistemi d’arma, in base alle tecnologie emergenti nel campo dei sistemi d’arma autonomi letali, che sono di interesse centrale per il Gruppo; (b) identificare e raggiungere una comprensione comune fra gli High Contracting Parties sui concetti e le caratteristiche dei sistemi d’arma autonomi potrebbe essere di aiuto per una ulteriore considerazione degli aspetti legati alle tecnologie emergenti nel campo dei LAWS […]; (c) differenti caratteristiche potenziali delle tecnologie emergenti nel campo dei sistemi d’arma autonomi letali, fra cui: autoadattamento; predicibilità; *explainability*; affidabilità; possibilità di essere oggetto di intervento; capacità di ridefinire o modificare obiettivi o traguardi o altrimenti adattarsi all’ambiente; e capacità di autoavviamento” (UN GGE CCW, 2019, p. 5).

In questo capitolo, nel paragrafo 6.2, introduco l’analisi comparativa delle definizioni esistenti degli AWS, con l’obiettivo di identificare i diversi approcci che ne sono alla base, i punti di somiglianza e di differenza, nonché i loro limiti. Nel paragrafo 6.3, analizzo gli aspetti essenziali degli AWS (autonomia, capacità di apprendimento, controllo umano, scopo d’uso) e propongo una definizione che costituisce un punto di partenza neutro rispetto ai valori etici e agli interessi politici. L’ambizione è che questa definizione possa favorire iniziative per la governance degli AWS e l’emergere di consenso a loro supporto. Trarrò le conclusioni nel paragrafo 6.4.

Prima di procedere con l’analisi, devo chiarire che, ai fini di questo capitolo, mi concentrerò sugli AWS e considererò i LAWS un sottoinsieme di quella categoria: in altre parole, agli scopi d’uso degli AWS – per esempio, anti-materiale, danno e distruzione – nel caso dei LAWS si aggiunge quello

dell’applicazione della forza letale. È un aspetto importante, perché i problemi etici legati agli AWS, come quelli di controllo, responsabilità e predicibilità, valgono a fortiori quando si prendono in considerazione i LAWS. Allo stesso tempo, però, questi ultimi pongono problemi etici specifici legati allo scopo letale del loro uso, per esempio per quanto riguarda la dignità umana e la virtù militare.

6.2 DEFINIZIONI DI SISTEMI D'ARMA AUTONOMI

La Tabella 6.2 elenca dodici definizioni di AWS o LAWS formulate da Stati e organizzazioni internazionali.[2]

Tabella 6.2 Dodici definizioni di AWS o LAWS presentate da Stati o organizzazioni internazionali fra 2012 e 2020.

Stato/organizzazione	Data	Definizione
Canada	2018	"Sistemi con la capacità di comporre in modo indipendente e di selezionare fra varie condotte per raggiungere obiettivi in base alle sue [informazioni] e alla conoscenza del mondo, di sé e della situazione." Nota: il Canada non ha una definizione ufficiale; questa è la definizione utilizzata dal Department of National Defence (Department of National Defence, 2018).
Cina	2018	"I LAWS devono includere, ma non esclusivamente, le seguenti 5 caratteristiche di base. La prima è la letalità, che significa un carico utile sufficiente e letale. La seconda è l'autonomia, che significa assenza di intervento e controllo umani durante tutto il processo di esecuzione di un compito. La terza è l'impossibilità di terminazione, il che significa che, una volta che il dispositivo si è avviato, non esiste modo di bloccarlo. La quarta è l'effetto indiscriminato, il che significa che il dispositivo porterà a termine il compito di uccidere e danneggiare indipendentemente da condizioni, scenari e bersagli. La quinta è l'evoluzione, nel senso che attraverso l'interazione con l'ambiente il dispositivo può apprendere in modo autonomo, ampliare le proprie funzioni e capacità in un modo che supera le aspettative umane" (Cina, 2018, p. 1). Nota: questa definizione è diversa da quella proposta dall'Esercito di Liberazione Popolare nel 2011: "[I LAWS sono] armi che utilizzano l'IA per perseguire, distinguere e distruggere automaticamente bersagli nemici; spesso comprendono sistemi di raccolta e gestione delle informazioni, sistemi di base di conoscenza, sistemi di assistenza alle decisioni, sistemi di implementazione della missione ecc." (Kania, 2018b).
Francia	2016	"Le armi autonome letali sono sistemi pienamente autonomi. Sono sistemi futuri: al momento non esistono. […] La definizione di LAWS implica una totale assenza di supervisione umana, nel senso che non esiste assolutamente alcun collegamento (comunicazione e controllo) con la catena di

Stato/organizzazione	Data	Definizione
		comando militare. […] La piattaforma di lancio di un LAWS deve essere in grado di spostarsi, adattarsi agli ambienti terrestri, marini o aerei e di individuare un bersaglio e attivare un effettore letale (pallottola, missile, bomba ecc.) senza alcun tipo di intervento o validazione umani. […] I LAWS con tutta probabilità avranno capacità di autoapprendimento" (République Française, 2016, pp. 1-2). "Data la complessità e varietà di ambienti (in particolare in aree urbane) e la difficoltà di creare algoritmi valoriali in grado di aderire ai principi del diritto umanitario internazionale (IHL), un LAWS con tutta probabilità possiede capacità di autoapprendimento, dato che sembra non realistico preprogrammare tutti gli scenari di un'operazione militare. Questo significa, per esempio, che il sistema di lancio deve essere in grado di selezionare un bersaglio indipendentemente dai criteri che sono stati predefiniti durante la fase di programmazione, in pieno rispetto dei requisiti dell'IHL. In base a quello che sappiamo delle future capacità tecnologiche, un LAWS pertanto sarebbe impredicibile" (*Ibidem*, p. 2).
Germania	2020	"I LAWS [sono] sistemi d'arma che escludono completamente il fattore umano da decisioni sul loro uso. Le tecnologie emergenti nel campo dei LAWS devono essere distinte, sul piano concettuale, dai LAWS. Le tecnologie emergenti, come digitalizzazione, intelligenza artificiale e autonomia, sono parte integrante dei LAWS, ma possono essere utilizzate nel pieno rispetto della legge internazionale" (Federal Foreign Office, 2020, p. 1).
International Committee of the Red Cross (ICRC)	2016	"Qualunque sistema d'arma con autonomia nelle proprie funzioni critiche. Ovvero, un sistema d'arma che possa selezionare (cioè, cercare o scoprire, identificare, tracciare, selezionare) e attaccare (cioè utilizzare la forza contro, neutralizzare, danneggiare o distruggere) bersagli senza intervento umano" (International Committee of the Red Cross, 2016, p. 1).
Israele	2018	"Nella prospettiva di Israele, punto di partenza condiviso per questa discussione deve essere che tutte le armi, inclusi i LAWS, sono e sempre saranno utilizzati da esseri umani. Dobbiamo evitare le visioni immaginarie in cui le macchine sviluppano, creano o attivano sé stesse – sono idee che vanno lasciate ai film di fantascienza. Per quanto riguarda la terminologia, questo significa che non bisogna pensare che i LAWS 'decidano' qualcosa. Sono sempre esseri umani quelli che decidono, e i LAWS sono soggetti alle loro decisioni" (Yaron, 2018, p. 2).

Stato/organizzazione	Data	Definizione
NATO		“Sistema automatizzato: un sistema che, in risposta agli input, segue un insieme predeterminato di regole e produce un esito predicibile.” “Sistema autonomo: un sistema che decide e agisce in modo da realizzare obiettivi desiderati, entro parametri ben definiti, in base alla conoscenza acquisita e a una *situational awareness* in evoluzione, seguendo una condotta ottimale ma potenzialmente impredicibile” (NATO, 2020, p. 16).
Norvegia	2017	“La Norvegia non ha ancora raggiunto una conclusione in merito a una definizione giuridica specifica del termine ‘sistemi d’arma completamente autonomi’. In termini generali, però, quando utilizziamo questo termine, ci riferiamo ad armi che cercano, identificano e attaccano bersagli, esseri umani compresi, utilizzando una forza letale senza che intervenga alcun operatore umano. Devono essere distinti da sistemi d’arma già in uso che sono altamente automatici, ma che operano entro limiti spaziali e temporali limitati tanto strettamente da ricadere al di fuori della categoria delle ‘armi completamente autonome’” (Norvegia, 2017, p. 1).
Svizzera		“Sistemi d’arma in grado di svolgere compiti governati dall’IHL con sostituzione parziale o completa di un essere umano nell’uso della forza, in particolare nel ciclo del bersaglio” (Svizzera, 2016, p. 2).
Paesi Bassi	2017	“Un’arma che, senza intervento umano, seleziona e ingaggia bersagli che soddisfano determinati criteri predefiniti, a seguito di una decisione umana di usare l’arma con l’idea che un attacco, una volta lanciato, non può più essere interrotto da un intervento umano” (Paesi Bassi, 2017, p. 1).
Regno Unito*	2018	“Un sistema autonomo è in grado di comprendere intenzioni e direttive di livello superiore. Da questa comprensione e dalla sua percezione dell’ambiente, un tale sistema è in grado di intraprendere un’azione appropriata per produrre uno stato desiderato. È in grado di decidere una linea di condotta, fra molte alternative, senza dipendere da supervisione e controllo umani, sebbene questi possano essere ancora presenti. Anche se l’attività complessiva di un velivolo autonomo senza pilota sarà prevedibile, le singole azioni potrebbero non esserlo” (Ministry of Defence, 2018a, p. 13).
	2016	“Il Regno Unito intende un tale sistema [LAWS completamente autonomo] come uno in grado di comprendere, interpretare e applicare intenzioni e direttive di livello superiore in base a una precisa comprensione e all’apprezzamento di quello che un comandante intende fare e, cosa forse ancora più importante, del perché. […] In maniera determinante, questa comprensione

Stato/organizzazione	Data	Definizione
		è focalizzata sull'effetto generale che l'uso della forza deve avere e sulla situazione desiderata che mira a produrre. Da questa comprensione, così come da una percezione sofisticata del suo ambiente e del contesto in cui opera, un sistema del genere deciderebbe di intraprendere (o di abortire) azioni appropriate per ottenere uno stato finale desiderato, senza supervisione umana, anche se un umano può essere comunque presente. L'output di un sistema del genere potrebbe, a volte, essere impredicibile – non seguirebbe semplicemente uno schema di regole con parametri definiti" (Foreign & Commonwealth Office, 2016, p. 2).
US Department of Defense	2012	"Un sistema d'arma che, una volta attivato, può selezionare e ingaggiare bersagli senza ulteriore intervento di un operatore umano. Questo include sistemi d'arma autonomi con supervisione umana che sono progettati in modo da consentire agli operatori umani di interrompere il funzionamento del sistema d'arma, ma che, dopo l'attivazione, possono selezionare e ingaggiare bersagli senza ulteriore input umano" (US Department of Defense, 2012, pp. 13-14).

* Il Regno Unito ha adottato la definizione di sistemi autonomi della NATO, ma non ha abbandonato alcuna delle proprie definizioni precedenti.

Questa pletora di definizioni ostacola il dibattito internazionale sulle implicazioni etiche e giuridiche degli AWS. Per esempio, secondo il report di Human Rights Watch,[3] ad agosto del 2020 trenta Stati avevano dichiarato di aderire a un bando preventivo degli AWS. Senza una definizione condivisa di AWS, però, è difficile identificare quali sistemi debbano essere messi al bando, peggio ancora far rispettare un simile bando. La Cina è un buon esempio in proposito. Roberts e colleghi (2020) evidenziano che i militari cinesi hanno espresso preoccupazione per l'uso dell'IA a scopi cinetici e aggressivi, e che quelle preoccupazioni hanno motivato il sostegno cinese a una restrizione dell'uso degli AWS, come espressa dalla Fifth Convention on CCW, e più di recente alla messa al bando dei LAWS. La definizione di "autonomia" adottata dalla Cina, però, è estremamente ristretta, poiché prende in considerazione solo armi *completamente* autonome e non tiene conto di AWS che possono avere livelli inferiori di autonomia (Kania, 2018b).

Anche altre definizioni si concentrano sull'autonomia completa. Quella del Regno Unito parla di sistemi completamente autonomi "in grado di

comprendere intenzioni e direttive di livello superiore". Il Regno Unito si concentra principalmente sull'intenzione del sistema, mentre i suoi partner internazionali si concentrano sull'intervento (o non intervento) umano nel sistema (Select Committee on Artificial Intelligence, 2018, p. 105). Questo punto è stato sostenuto in varie riunioni dell'UN GGE e in un report dello House of Lords' Select Committee on Artificial Intelligence.[4] La definizione fa riferimento a capacità cognitive che i sistemi IA attualmente non possiedono e che è poco probabile acquisiscano in futuro (Floridi, 2014; Wooldridge, 2020). Riferirsi ad AWS "in grado di comprendere intenzioni e direttive di livello superiore" determina una soglia molto alta, inverosimile da raggiungere, per ciò che va considerato autonomo. La definizione della Francia è analoga, e ribadisce esplicitamente che in base alla sua definizione gli AWS "al momento non esistono".

L'approccio della Francia ha anche l'effetto di orientare le direzioni future dell'innovazione tecnologica, indicando dei limiti ai possibili usi delle tecnologie IA. In tal modo gli enti di regolamentazione possono avere un vantaggio sull'innovazione. Questa impostazione però si basa su una concezione paternalistica del ruolo delle regolamentazioni e dei relativi enti, che è in sé problematica e può avere l'effetto indesiderato di ostacolare l'innovazione. Pensando specificamente agli AWS, definire la governance di questi sistemi concentrandosi su scenari futuristici è pericoloso per due motivi. Il primo è che, se ci si concentra su sistemi non ancora sviluppati o le cui caratteristiche sono irrealizzabili sul piano tecnologico, si distoglie l'attenzione da urgenti problemi etici e giuridici posti dagli AWS già esistenti e da quelli che potranno essere utilizzati in un futuro prevedibile. Il secondo motivo è che si minano alla base regolamentazioni e dichiarazioni sul bando degli AWS, se queste fanno riferimento ad AWS ipotetici con caratteristiche che i sistemi attuali e quelli prevedibili non possiedono – per esempio, la capacità di comprensione e intenzione. In questo caso, l'implicazione è che le dichiarazioni ufficiali sul bando degli AWS fanno riferimento a sistemi che non esistono ancora e ignorano invece altri sistemi attualmente in uso.

Lo stesso vale per il Regno Unito. L'ONG Article36 sottolinea che affermazioni fatte dal Regno Unito come, per esempio, "non abbiamo in programma di sviluppare o acquisire armi del genere" – in base alla definizione britannica di AWS – "possono apparire progressiste, senza di

fatto porre alcun vincolo alla capacità del Regno Unito di sviluppare sistemi d'arma con autonomia sempre maggiore" (Article36, 2018, p. 1). In effetti, la soglia molto elevata fissata per identificare gli AWS, se rimarrà immutata, permetterà al Regno Unito di utilizzare AWS a meno che questi non presentino una "comprensione di intenzioni e direttive di livello superiore" (vedi la Tabella 6.2). Il problema in questo caso è concettuale: la definizione restrittiva di AWS non consente la classificazione corretta di tali sistemi, che sono autonomi, ma non superano la soglia stabilita dalla definizione britannica. Questi sistemi o ricadono in un'area grigia tra le categorie o vengono erroneamente inseriti nella categoria più familiare dei sistemi automatici, e così si evita di considerare e affrontare i problemi etici e giuridici che sollevano.

Per non incorrere in questi limiti, è importante definire gli AWS concentrandosi sui loro aspetti caratterizzanti (per esempio, l'autonomia) e descriverli in base a una comprensione scientifica e tecnologica. In tal modo la definizione può offrire uno strumento rigoroso per identificare gli AWS e per evitare di concentrarsi su caratteristiche che questi sistemi non hanno o che potrebbero non esibire mai. L'obiettivo di una simile definizione, come afferma l'ICRC, è quello di

> includere alcuni sistemi d'arma esistenti, [e così] rendere possibile una considerazione realistica della tecnologia delle armi, per valutare che cosa possa rendere accettabili (sul piano giuridico ed etico) certi sistemi d'arma esistenti e quali sviluppi tecnologici emergenti possano far sorgere preoccupazioni in base al diritto umanitario internazionale (IHL), ai principi dell'umanità e a ciò che impone la coscienza pubblica. (International Committee of the Red Cross, 2016, p. 1)

Questo è, per esempio, il fondamento della definizione dell'ICRC (vedi la Tabella 6.2) e l'esito della definizione degli Stati Uniti, che considera l'autonomia uno spettro basato sulla funzione rispetto al coinvolgimento umano, in modo da poter includere anche i sistemi d'arma *esistenti* (International Committee of the Red Cross, 2016, p. 1; US Department of Defense, 2012, pp. 13-14).

Per cercare di arrivare a una definizione inclusiva, però, è importante anche mantenere un certo livello di specificità, per evitare un approccio troppo generico, che potrebbe generare confusione. Questo è il rischio che corre la definizione della NATO (vedi la Tabella 6.2). La definizione non vuole fissarsi specificamente sugli AWS bensì sui *sistemi autonomi* in generale, tuttavia finisce per risultare troppo generica. Per esempio, parla

degli “obiettivi desiderati” di un sistema, senza specificare se si tratti di obiettivi politici, organizzativi, strategici o tattici oppure degli obiettivi specifici che il sistema stesso può avere o acquisire. Analogamente, parla di *situational awareness*, ma non è chiaro se con questa espressione si intenda una comprensione del contesto immediato di uso del sistema o dello scenario strategico più ampio.

Dall’analisi delle definizioni riportate nella Tabella 6.2, si possono ricavare quattro caratteristiche che vengono citate più spesso nella definizione di AWS, cioè autonomia, capacità di apprendimento, intervento e controllo umano, e scopo d’uso. Queste caratteristiche vanno nella direzione giusta nel considerare ciò che gli AWS sono – sono in accordo, per esempio, con la definizione di IA adottata da Taddeo, McCutcheon e Floridi (2019) e da Taddeo e colleghi (2021) come una forma di agency autonoma e capace di autoapprendimento. I prossimi tre paragrafi analizzano separatamente queste caratteristiche, chiarendone le implicazioni per quanto riguarda il dibattito etico sugli AWS.

6.2.1 Autonomia, intervento e controllo

L’autonomia è un elemento centrale di tutte le definizioni di AWS elencate nella Tabella 6.2. In qualche caso viene considerata come la capacità di un sistema di operare con successo senza intervento umano. La definizione proposta dalla Germania, per esempio, parla di macchine che “escludono completamente” gli esseri umani dai processi decisionali. In altri casi, l’autonomia è fusa con l’assenza di controllo umano. Così è, per esempio, nella definizione francese, per cui i LAWS non hanno “assolutamente alcun collegamento (comunicazione e controllo) con la catena di comando militare” (République Française, 2016, p. 1).

Come vedremo nel paragrafo 6.3, mettere insieme autonomia e assenza di controllo umano è fuorviante, sul piano concettuale come su quello operativo. Un sistema IA pienamente autonomo può comunque essere utilizzato sotto qualche forma di controllo umano.

La distinzione fra autonomia e controllo è importante perché offre tre vantaggi. Il primo: per chiarezza concettuale, automazione e controllo umano non sono concetti mutuamente esclusivi. L’automazione rende non necessario l’intervento umano per il raggiungimento di un dato obiettivo, ma questo non esclude qualche forma di controllo umano, per esempio il

monitoraggio del comportamento di un sistema per disattivarlo se fa qualcosa di sbagliato. Per questo la Direttiva 3000.09 del DoD è corretta, quando fa esplicito riferimento a "sistemi d'arma autonomi con supervisione umana" (US Department of Defense, 2012, p. 14) e li distingue dai "sistemi d'arma semi-autonomi", la cui autonomia è circoscritta alle "funzioni relative all'ingaggio", che dipendono da un operatore umano per la selezione del bersaglio.

Un secondo vantaggio della distinzione fra autonomia e controllo è che mantiene aperti futuri sviluppi del dibattito sugli AWS. Molti dei problemi non riguardano il livello di autonomia desiderabile di questi sistemi, bensì il livello desiderabile di controllo umano su di essi. La decisione in merito al controllo è per molti versi normativa, in quanto non è definita solo dalle *affordances* tecnologiche (per esempio, dalla rapidità di un sistema), ma anche, cosa ancora più importante, dalle decisioni e dai compiti che possono essere delegati alle macchine senza richiedere un controllo umano. La separazione dei due concetti consente di concentrarsi sulle forme di controllo desiderabili sul piano normativo, indipendentemente dal livello di autonomia che quelle macchine potrebbero acquisire in futuro.

Il terzo vantaggio della distinzione fra autonomia e controllo è che anticipa gli approcci che si servono della mancanza di esempi concreti di AWS pienamente autonomi per concludere che non esistono problemi in merito al controllo degli AWS, dato che gli esseri umani sono ancora coinvolti (o *on the loop*). Questo è l'atteggiamento che spesso adotta chi tenta di evitare di discutere della necessità di regolamentare/bandire l'uso degli AWS (Federazione Russa, 2017, p. 2).

6.2.2 Capacità di apprendimento

Delle dodici definizioni prese in considerazione in questo capitolo, solo quella francese e quella cinese evidenziano le capacità di apprendimento come una caratteristica fondamentale degli AWS. La mancata considerazione delle capacità di apprendimento nella definizione degli AWS è problematica, perché questo è un aspetto fondamentale delle tecnologie IA. Ovviamente, gli AWS possono funzionare senza capacità di apprendimento: per esempio, una programmazione basata su regole consente una reazione immediata a segnali ambientali, anche se non la

pianificazione di nuovi comportamenti quando le condizioni ambientali cambiano. Si può immaginare che un sensore rilevi l'imminente arrivo di un oggetto e che l'algoritmo attivi una risposta del sistema – per esempio, sparare per distruggere quell'oggetto.

Tuttavia, con gli sviluppi della tecnologia i sistemi ad algoritmi basati su regole vengono progressivamente sostituiti da sistemi basati su IA. Le istituzioni militari investono in IA per un insieme molto ampio di applicazioni. Per esempio, sono già in corso iniziative significative per sfruttare gli sviluppi nel riconoscimento di immagini, volti e comportamenti mediante tecniche IA e ML ai fini di raccolta di intelligence e "riconoscimento automatico di bersagli" per identificare persone, oggetti o schemi comportamentali (abbiamo parlato di IA per l'analisi dell'intelligence nel capitolo 3). La definizione francese di AWS sottolinea che le capacità di apprendimento sono necessarie per adattarsi alla complessità degli scenari operativi che non possono essere previsti e quindi non possono essere "preprogrammati" nel sistema.

Se non si tiene conto delle capacità di apprendimento nelle definizioni di AWS, se ne trascura una caratteristica chiave, uno degli aspetti principali che ne guida l'adozione, e si ostacola la discussione sulle loro implicazioni etiche e giuridiche. Come abbiamo visto nel capitolo 1, le capacità di apprendimento dell'IA suscitano interrogativi a causa della loro predicibilità, e quindi della loro affidabilità; rispetto all'attribuzione della responsabilità delle azioni che questi sistemi compiono; e anche per quanto riguarda l'implementazione di forme significative di controllo. Come riportato nella definizione francese:

> In base a quello che sappiamo delle future capacità tecnologiche, un LAWS pertanto sarebbe *impredicibile*. (République Française, 2016, p. 2, corsivo mio)

L'ICRC mette in evidenza un punto simile: "L'applicazione di IA e ML a funzioni di scelta del bersaglio solleva questioni fondamentali di impredicibilità intrinseca" (International Committee of the Red Cross, 2018, p. 2). Le capacità di apprendimento (che portano all'impredicibilità degli esiti) sollevano problemi rispetto all'Articolo 36 del Protocollo aggiuntivo I alle Convenzioni di Ginevra sulle nuove armi:

> Da un punto di vista tecnico, qualsiasi sistema che continui ad apprendere quando è in uso cambia costantemente. Non è lo stesso sistema che era quando è stato introdotto o verificato per l'introduzione. Sono stati sollevati interrogativi sulla legalità di sistemi

adattivi, in particolare in riferimento agli obblighi degli Stati previsti dall'Articolo 36. (UNIDIR, 2017, p. 10)

Questo è fondamentale, come nota l'ICRC:

> La possibilità di eseguire una verifica [in base all'Articolo 36] comporta una comprensione piena delle capacità delle armi e la previsione dei loro effetti, in particolare grazie a test. Prevedere tali effetti però può diventare sempre più difficile, se i sistemi d'arma autonomi dovessero diventare più complessi o avere una maggiore libertà d'azione nelle loro operazioni, e pertanto diventassero meno predicibili. (Riportato in UNIDIR, 2017, p. 26)

Per ragioni sia etiche sia giuridiche, pertanto, tenere conto delle capacità di apprendimento degli AWS è essenziale. È la natura del processo di apprendimento che crea al tempo stesso opportunità e sfide significative e distingue nettamente i sistemi abilitati dall'IA da quelli altamente automatizzati ma basati su regole predefinite. Le capacità di apprendimento caratterizzano le generazioni più recenti, e quelle future, degli AWS. Tenerle ben presenti consente un ulteriore chiarimento della distinzione fra sistemi automatici e autonomi (riprenderemo il tema nel paragrafo 6.3) e permette di identificare la fonte di molte implicazioni etiche e giuridiche fondamentali degli AWS. Per questo è importante che le definizioni degli AWS citino esplicitamente quelle capacità. È problematico che perfino le due definizioni più esaustive (quelle degli USA e dell'ICRC) non ne tengano conto, lasciandosi sfuggire l'opportunità di fare luce su un elemento chiave di questi sistemi.

6.2.3 Scopo d'uso

La maggior parte delle definizioni nella Tabella 6.2 definisce lo scopo d'uso degli AWS solo implicitamente, parlando di "armi" e dicendo che gli AWS sono utilizzati in contesti cinetici. Ciò indica qualche forma di uso distruttivo (anti-materiale o letale) di questi sistemi. È importante però comprendere la gamma dei possibili usi con una maggior precisione, per esempio considerando i compiti specifici che gli AWS possono svolgere nel contesto di operazioni cinetiche.

Fra le definizioni nella Tabella 6.2, quattro (quelle di Canada, Israele, Germania e Regno Unito) non citano esplicitamente alcuno scopo d'uso specifico. In tal caso l'esito cinetico dell'uso degli AWS è dato per scontato, lasciando non definito, per esempio, se verranno utilizzati per un

bersaglio deliberato o dinamico. Anche la definizione della NATO non cita alcuno scopo specifico (va sottolineato, però, che la definizione della NATO è riferita a sistemi autonomi in generale e non agli AWS). Le altre parlano di AWS che esercitano una forza letale (Cina e Francia) o più specificamente selezionano o colpiscono bersagli (umani o no) da neutralizzare, danneggiare o distruggere (ICRC, Norvegia, Svizzera, Paesi Bassi, Stati Uniti).

Nessuna delle definizioni considera i passaggi specifici che una macchina deve compiere per eseguire i compiti a essa delegati. Eppure, proprio tali passaggi risultano centrali quando si analizzano gli AWS e le implicazioni etiche e giuridiche connesse al loro impiego. Si consideri, in tal senso, la critica mossa da Roff (2014) alla definizione adottata dagli Stati Uniti: il termine "selezionare", nell'espressione "selezionare e ingaggiare", appare ambiguo, poiché non è chiaro se comprenda anche la fase di individuazione del bersaglio (Conn, 2016). Qualora tale individuazione non fosse ricompresa nella definizione, si potrebbe presumere che essa sia affidata a un operatore umano – ipotesi che eluderebbe molte delle questioni etiche (e tecniche) più controverse.

La critica avanzata da Roff mette in luce la complessità intrinseca di tali compiti e dei processi che sorreggono la decisione di ricorrere all'uso della forza. Si consideri, per esempio, la sequenza di passaggi che compongono il processo decisionale in materia di individuazione degli obiettivi, come descritto in Ekelhof e Persi Paoli (2020). Gli autori delineano un meccanismo articolato, che si sviluppa lungo l'intera catena decisionale e di comando quando si prende in esame l'impiego degli AWS. Tale processo comprende attività e scelte che spaziano dai livelli tattico e operativo fino a quelli strategico e politico, i quali risultano spesso intrecciati in modo profondo. La complessità del processo richiede un approccio più specifico quando si considerano i compiti svolti da un AWS, che si articola in tre modi, precisando esplicitamente:

- le finalità per cui tali sistemi vengono impiegati e gli obiettivi distruttivi perseguiti, siano essi letali o no;
- i passaggi del processo di applicazione della forza che rientrano nell'ambito operativo del sistema autonomo;
- il grado di controllo umano sotto il quale esso è chiamato ad agire.

Sono queste specificazioni a determinare l’esito delle analisi etiche e giuridiche relative agli AWS.

6.3 UNA DEFINIZIONE DI AWS

Nel proporre una nuova definizione di AWS, il mio obiettivo è duplice: chiarire le caratteristiche fondamentali che permettono l'identificazione degli AWS, e specificarne relazioni (per esempio, automazione e controllo) e differenze (per esempio, automatico rispetto ad autonomo). Per farlo, considero autonomia, capacità di apprendimento e controllo come caratteristiche che si collocano su un continuum: gli AWS possono avere ciascuna di quelle caratteristiche in misura maggiore o minore. Per quanto poi riguarda l'insieme dei possibili scopi d'uso, adotto un approccio inclusivo, con l'obiettivo di chiarire quale possa essere la loro estensione. Identificare la combinazione dei diversi livelli di queste caratteristiche e degli (eventuali) scopi d'uso che possono soddisfare particolari requisiti etici e giuridici è il compito di analisi etiche (che saranno sviluppate nei prossimi capitoli). Tenendo presente tutto questo, ecco la mia definizione:

> Un AWS è un agente artificiale che, come minimo, è in grado di modificare i propri stati interni per raggiungere uno o più obiettivi dati all'interno del proprio ambiente operativo dinamico e senza l'intervento diretto di un altro agente (è, cioè, un agente artificiale automatizzato) e può essere anche in grado di modificare le proprie regole di transizione per adattarle all'ambiente o per perfezionare il proprio comportamento (ha, cioè, capacità di apprendimento) senza l'intervento diretto di un altro agente, ed è utilizzato per esercitare una forza cinetica nei confronti di un oggetto fisico o di un essere umano e, a questo fine, è in grado di identificare, selezionare e attaccare il bersaglio senza l'intervento diretto di un altro agente. Una volta in uso, un AWS può operare con o senza controllo umano.

Nei prossimi paragrafi approfondisco questa definizione concentrandomi sui concetti di autonomia, capacità di apprendimento e controllo. Gli scopi d'uso sono meno problematici dal punto di vista concettuale, perciò non li analizzerò ulteriormente. È importante, però, dire qui che gli scopi d'uso identificati sono quelli direttamente relativi all'obiettivo, e riguardano l'esercizio della forza. La selezione dei bersagli e il fatto di colpirli (in modo deliberato oppure dinamico) sono collegati direttamente all'uso della forza. Quindi, un sistema le cui funzioni di selezione e attacco sono autonome, ma che è diretto da altri agenti per tutti i suoi altri scopi d'uso, per esempio la mobilità, sarebbe considerato ancora un AWS.

6.3.1 Sistemi d'arma autonomi ad autoapprendimento

Un tema fondamentale alla base della definizione di AWS è la distinzione fra sistemi *automatici*, *automatizzati* e *autonomi*. In particolare, la distinzione fra automatico e autonomo può essere difficile da fare quando la si considera a un LdA etico. Un report dell'ICRC, per esempio, sottolinea che "non esiste una distinzione tecnica chiara fra sistemi automatizzati e autonomi, e non esiste un accordo universale sul significato di questi termini" (International Committee of the Red Cross, 2019, p. 7). In modo analogo, la Joint Concept Note 1/18, "Human-machine teaming", pubblicata dal ministero della Difesa del Regno Unito nel 2018, inizia osservando che "non esiste un confine chiaro, definibile e universalmente condiviso fra ciò che costituisce automazione e ciò che è autonomo […] perché la valutazione di autonomia e l'uso del termine sono soggettivi e contestuali" (Ministry of Defence, 2018b, p. 57). Si può essere d'accordo che la distinzione fra automazione e autonomia sia sfumata, ma non è tale perché la valutazione dell'autonomia degli agenti artificiali sia soggettiva o dipendente dal contesto. Nel campo della computer science, e in particolare della *teoria degli agenti*, le differenze fra questi concetti sono chiare (Wooldridge, Jennings, 1995; Castelfranchi, Falcone, 2003).

Prendiamo in considerazione per primi gli agenti automatici. Questi sono agenti le cui azioni sono predeterminate e non cambiano, a meno che si presentino fattori di attivazione preselezionati o vi sia un intervento umano. Gli agenti automatici non sono teleologici; non perseguono un obiettivo, ma semplicemente reagiscono a uno stimolo esterno. In tal senso sono "entità causali" (Castelfranchi, Falcone, 2003). Una mina rientra esattamente in questa categoria, perché la sua azione è determinata in modo causale da un evento specifico, per esempio la pressione esercitata da qualcuno che ci cammini sopra. Gli AWS non appartengono a questa categoria, nella misura in cui il loro comportamento non è predeterminato.

Gli AWS eseguono dei compiti per raggiungere degli obiettivi (agenti teleologici); possono regolare le proprie azioni sulla base del feedback che ricevono dall'ambiente (agenti automatizzati), possono essere in grado di definire dei piani (agenti euristici) per raggiungere il loro obiettivo, e possono anche essere in grado di modificare il proprio comportamento in risposta a cambiamenti che avvengono nell'ambiente (agenti ad

apprendimento autonomo). A questo punto, possiamo considerare gli AWS come sistemi che, come minimo, sono agenti artificiali teleologici automatizzati, ma possiamo anche essere più specifici e fare un passo ulteriore.

Ai fini della definizione, è importante considerare quali siano i requisiti minimi che deve soddisfare un agente artificiale per essere autonomo. Per questo facciamo riferimento alle definizioni di agenti artificiali autonomi offerte da Castelfranchi e Falcone (2003) e da Floridi e Sanders (2004).

Secondo Castelfranchi e Falcone, gli agenti autonomi hanno le seguenti proprietà:

> Il loro comportamento è *teleonomico*: tende a certi risultati specifici in conseguenza di vincoli o rappresentazioni interne, prodotti da progettazione, evoluzione o apprendimento […]; non si limitano semplicemente a ricevere un input – non semplicemente una forza (energia) ma informazione – ma *"percepiscono" e interpretano (attivamente) il loro ambiente* e gli effetti delle loro azioni; […] *si orientano verso l'input*; in altre parole, definiscono e selezionano gli stimoli ambientali; […] *hanno "stati interni"*, con i propri principi di evoluzione esogeni ed endogeni, e anche il loro comportamento dipende da tali stati interni. (2003, p. 105)

Lo stato interno di un agente artificiale può essere descritto come la configurazione dell'agente (per esempio, i valori e i pesi di una rete neurale in un istante specifico) quando svolge una data operazione. Gli stati interni sono fondamentali nella definizione di autonomia, e la transizione fra stati corrisponde a una variazione nel comportamento del sistema. Il modo in cui è determinata la transizione definisce la differenza tra sistemi automatizzati e autonomi. Gli stati interni sono fondamentali anche per la definizione offerta da Floridi e Sanders, in cui gli agenti artificiali autonomi presentano tre caratteristiche:

> *Interattività* significa che l'agente e il suo ambiente (possono) agire l'uno sull'altro. Esempi tipici sono l'input o l'output di un valore, o il simultaneo manifestarsi di un'azione in agente e paziente – per esempio la forza gravitazionale fra corpi.
>
> *Autonomia* significa che l'agente è in grado di modificare il proprio stato senza che questo costituisca una risposta diretta all'interazione: può effettuare transizioni interne per modificare il proprio stato […].
>
> *Adattabilità* significa che le interazioni dell'agente cambiano (possono cambiare) le regole di transizione in base alle quali cambia il suo stato. Questa proprietà garantisce che un agente, a un dato LdA, possa essere considerato in grado di apprendere la propria modalità di funzionamento in un modo che dipende decisamente dalla sua esperienza. (2004, p. 357)

Il fatto che un agente artificiale possa modificare i propri stati interni senza l'intervento diretto di un altro agente costituisce la linea di confine tra automatico/automatizzato e autonomo. In base a questo criterio, sono autonomi sia un sistema artificiale basato su regole, sia un sistema che apprende.

Come si è già detto nel paragrafo 6.2.1, la capacità di apprendimento è una caratteristica sempre più comune degli AWS. È la caratteristica che ne determina la capacità di affrontare scenari complessi e in rapida evoluzione, ma anche quella che porta all'impredicibilità, alla mancanza di trasparenza e di controllo e al *responsibility gap* in merito all'uso di questi agenti. È importante, perciò, includerla nella definizione di AWS, offrirne una specificazione chiara per evitare di antropomorfizzare questi agenti, e fissare una soglia precisa al di sotto della quale si può dire che un agente non ha capacità di apprendimento. Per questa ragione, nella definizione di AWS proposta qui, faccio riferimento a un agente artificiale dotato di qualche capacità di modificare *le proprie regole di transizione* per operare con successo in un ambiente che cambia.

6.3.2 Controllo umano

La definizione di AWS si riferisce al controllo umano come modalità d'uso degli AWS e non come una loro caratteristica definitoria. Esistono forme differenti di controllo (Tsamados, Taddeo, 2023). Amoroso e Tamburrini, per esempio, ne identificano tre:

> In primo luogo, l'obbligo di rispettare l'IHL comporta che il controllo umano deve avere un ruolo da attore "a prova di guasto", che contribuisce a impedire che il malfunzionamento dell'arma abbia come risultato un attacco diretto contro la popolazione civile o danni collaterali eccessivi. In secondo luogo, per evitare vuoti di *accountability*, il controllo umano deve fungere da attrattore di *accountability*, cioè deve garantire le condizioni *giuridiche* per l'attribuzione di responsabilità, nel caso in cui un'arma segua un tipo di condotta che viola la legge internazionale. In terzo luogo, dal principio del rispetto della dignità umana segue che il controllo umano deve operare come attivatore di agency morale, garantendo che le decisioni che influiscono sulla vita, l'integrità fisica e i beni delle persone (combattenti inclusi) coinvolte nei conflitti armati non vengano prese da agenti artificiali non morali. (2020, p. 189)

Si può non essere d'accordo con questa tassonomia, o pensare che sia meglio definire il controllo a un LdA diverso, per esempio, concentrandosi solo sulle specifiche tecniche degli AWS. La letteratura in materia, però, converge nel considerare il controllo degli AWS dinamico,

multidimensionale e dipendente dalla situazione, e come una prassi che può essere esercitata focalizzandosi su aspetti diversi del team umani-macchine. Per esempio, lo Stockholm International Peace Research Institute e l'ICRC identificano tre aspetti principali del controllo umano dei sistemi d'arma: i parametri d'uso del sistema, l'ambiente e l'interazione umani-macchine (Boulanin et al., 2020). Si possono prendere in considerazione anche altri aspetti. Secondo Boardman e Butcher (2019), il controllo non deve essere solo significativo ma anche appropriato, perché deve essere esercitato in modo da garantire che il coinvolgimento umano nel processo decisionale rimanga significativo, senza con questo compromettere le prestazioni del sistema.

La discussione su che cosa costituisca un controllo umano significativo degli AWS e se si possa esercitare in modo appropriato non rientra negli scopi di questo capitolo, che sono invece l'identificazione delle caratteristiche degli AWS, più che le condizioni normative per la loro progettazione, sviluppo e uso. Tuttavia, dato che questa analisi può gettare luce su queste caratteristiche e sulle loro relazioni, è importante sottolineare che il controllo umano non è antitetico all'autonomia degli AWS e può essere esercitato su questi sistemi a livelli diversi, dalle decisioni politiche e strategiche in merito all'uso degli AWS ai tipi di compiti che vengono delegati a questi sistemi. L'interrogativo è quale forma di controllo sia desiderabile dal punto di vista etico e, assumendo che sia fattibile, debba essere presa in considerazione dai decisori politici nella progettazione di un quadro generale per la governance degli AWS.

6.4 CONCLUSIONE

Il dibattito sui sistemi d'arma autonomi (AWS) è profondamente influenzato da considerazioni strategiche, politiche ed etiche. Interessi contrapposti e valori divergenti tendono ad accentuarne la polarizzazione, mentre definizioni cariche di implicazioni politiche minano i tentativi di delineare impieghi legittimi e di stabilire un quadro normativo adeguato. A complicare ulteriormente tali sforzi contribuisce la persistente confusione concettuale che circonda il tema.

In questo capitolo, era mia intenzione superare la confusione concettuale che circonda i sistemi d'arma autonomi (AWS), attraverso un'analisi comparativa delle definizioni attualmente in uso e la proposta di una nuova definizione, neutra sotto il profilo valoriale. L'esame comparato delle definizioni esistenti consente di individuare i principali nodi di ambiguità concettuale – fra cui la distinzione tra "automatico" e "autonomo", e quella, altrettanto cruciale, tra "autonomia" e "controllo". L'analisi mette inoltre in luce una lacuna rilevante: la scarsa attenzione riservata alla capacità di apprendimento di tali sistemi.

La definizione proposta, priva di riferimenti a obiettivi politici o strategici e scevra di elementi normativi, è costruita a partire dalle caratteristiche tecniche fondamentali degli AWS, con l'unico scopo di permettere la loro corretta identificazione e di distinguerli da altri sistemi d'arma, come quelli automatici. Ritengo che una definizione neutrale dal punto di vista valoriale possa contribuire in modo significativo al dibattito accademico e politico su questo tema, fornendo un terreno comune su cui far convergere posizioni differenti.

---

1. Vedi sulla necessità, Blanchard, Taddeo, 2022a, 2022b; sulla proporzionalità nell'applicazione agli AWS, Blanchard, Taddeo, 2022c.

2. La NATO offre una definizione di *sistemi autonomi* e non specificamente di AWS. La includo comunque, perché fa riferimento a caratteristiche identificative degli AWS.

3. Human Rights Watch, "Stopping killer robots", 10 agosto 2020, https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and.

4. Select Committee on Artificial Intelligence, 2018, p. 105. Il 24 aprile 2019 Lord Browne ha presentato un'interrogazione alla Camera dei Lord a proposito delle indicazioni ricevute dal ministero della Difesa in merito alla raccomandazione di allineare la definizione del Regno Unito degli AWS a quella dei partner internazionali. Il governo rispondeva di aver ricevuto alcune indicazioni, ma precisava comunque che "la UN Convention on Certain Conventional Weapons Group of Government Experts on Lethal Autonomous Weapons Systems deve ancora raggiungere un accordo su una definizione accettata a livello internazionale o su un insieme di caratteristiche per le armi autonome" (House of Lords, 2019).

7

MORAL GAMBIT

ACCETTARE LA RESPONSABILITÀ MORALE PER LE AZIONI DI SISTEMI D'ARMA AUTONOMI

7.1 INTRODUZIONE

Una questione fondamentale, quando si considera l'uso dell'IA nella difesa, è l'attribuzione della responsabilità morale per le azioni compiute da sistemi IA. La questione è pertinente per tutte le tre categorie di usi dell'IA nella difesa, ma diventa estremamente rilevante per gli usi conflittuali e cinetici, in particolare degli AWS, perché intenzioni, piani, diritti, doveri, elogi o punizioni si possono attribuire in modo sensato solo a esseri umani. Attribuire questa responsabilità agli AWS, o in generale ai sistemi IA, comporterebbe un'errata attribuzione di "*accountability* causale e responsabilità giuridica per quanto riguarda i loro errori e i loro abusi. Si potrebbe dare la colpa ai robot per punirli al posto degli esseri umani. E persone irresponsabili non riconoscerebbero la necessità di prestare attenzione all'ingegnerizzazione, al marketing e all'uso dei robot" (Floridi, Taddeo, 2018, p. 309).

Il consenso sull'attribuzione agli agenti umani della responsabilità morale per le azioni degli AWS si è allargato al punto che oggi questa posizione non è più controversa. Per esempio, la UN GGE CCW identifica la responsabilità umana come principio fondamentale per l'uso (possibile) dei LAWS, precisando che "la responsabilità umana per le decisioni in merito all'uso dei sistemi d'arma deve essere mantenuta, perché la *accountability* non può essere trasferita alle macchine. Questo va considerato su tutto il ciclo di vita del sistema d'arma" (UN GGE CCW, 2019).

Come si ricorderà dal capitolo 1, la responsabilità umana è citata esplicitamente nei principi etici per l'uso dell'IA nelle Recommendations on the Ethical Use of Artificial Intelligence del DoD degli Stati Uniti, il cui primo principio afferma che "gli esseri umani devono esercitare livelli appropriati di giudizio e rimanere responsabili dello sviluppo, della messa in campo, dell'uso e degli esiti dei sistemi IA del DoD" (Defense Innovation Board, 2019, p. 8). Analogamente, in un report per il Parlamento europeo, il Committee for Legal Affairs ha affermato che "i processi decisionali autonomi non devono assolvere gli esseri umani dalla responsabilità, e […] le persone devono avere sempre la responsabilità

ultima per i processi decisionali, in modo che l'essere umano responsabile della decisione possa venire identificato" (Lebreton, 2021).

Il consenso su questo punto riveste un'importanza cruciale, poiché consente di eludere il rischio di antropomorfizzare gli AWS – e, più in generale, l'IA – evitando altresì di disperdere energie nel dibattito, ampiamente trattato in letteratura, circa l'attribuzione di una responsabilità morale a tali sistemi. Dal momento che gli AWS non posseggono intenzionalità né alcuna comprensione del biasimo o dell'elogio che potrebbe conseguire alle loro azioni, non può esservi, in senso proprio, alcuna imputazione morale a loro carico.

Tuttavia, sebbene il focus sugli agenti umani indirizzi correttamente il dibattito, permane una difficoltà sostanziale nell'attribuire, in modo compiuto e significativo, una responsabilità morale per le azioni compiute dagli AWS – questione sulla quale torneremo fra breve. L'attribuzione di questa responsabilità agli agenti umani, invece che a entità astratte come istituzioni o soggetti giuridici, costituisce una condizione imprescindibile per l'impiego legittimo degli AWS. Come fu inequivocabilmente affermato negli atti dei processi di Norimberga seguiti alla Seconda guerra mondiale: "I crimini contro il diritto internazionale sono commessi da uomini, non da entità astratte; solo punendo gli individui che li compiono è possibile garantire l'effettiva applicazione del diritto internazionale" (International Military Tribunal, Nuremberg, 1947, p. 221). Nell'epoca della guerra autonoma, gli AWS possono compiere azioni moralmente riprovevoli; ma solo attribuendo responsabilità morale a coloro che li progettano, sviluppano e impiegano si potrà preservare una nozione eticamente coerente della condotta bellica. La questione è se e come questa responsabilità possa essere attribuita in modo appropriato.

In questo capitolo procederò a un'analisi delle principali posizioni emerse nel dibattito concernente la responsabilità morale in relazione ai sistemi di intelligenza artificiale in generale, esaminandole nei paragrafi 7.2 e 7.3. Successivamente mi concentrerò su approcci specifici volti ad attribuire responsabilità morale nel caso dei sistemi d'arma autonomi (AWS), valutandone punti di forza e criticità, nel paragrafo 7.4. Nel paragrafo 7.5 offrirò il mio personale contributo al dibattito, soffermandomi sui concetti di *responsabilità morale significativa* e di *scommessa morale*. Seguiranno, nel paragrafo 7.6, otto raccomandazioni mirate, destinate in particolare alle istituzioni della difesa, su come

colmare il cosiddetto *responsibility gap* nell'impiego di *AWS non letali*. Il capitolo si conclude con una sintesi finale nel paragrafo 7.7.

Prima di iniziare l'analisi sono necessari tre chiarimenti. Innanzitutto, in questo capitolo mi concentrerò sugli AWS in base alla definizione che ne ho dato nel capitolo 6:

> Un AWS è un agente artificiale che, come minimo, è in grado di modificare i propri stati interni per raggiungere uno o più obiettivi dati all'interno del proprio ambiente operativo dinamico e senza l'intervento diretto di un altro agente (è, cioè, un agente artificiale automatizzato) e può essere anche in grado di modificare le proprie regole di transizione per adattarle all'ambiente o per perfezionare il proprio comportamento (ha, cioè, capacità di apprendimento) senza l'intervento diretto di un altro agente, ed è utilizzato per esercitare una forza cinetica nei confronti di un oggetto fisico o di un essere umano e, a questo fine, è in grado di identificare, selezionare e attaccare il bersaglio senza l'intervento diretto di un altro agente. Una volta in uso, un AWS può operare con o senza controllo umano.

Questa è una definizione neutra rispetto ai valori, perciò nella nostra analisi non ha altra funzione che quella di identificare l'insieme degli AWS a cui si applica. Vale la pena di notare ancora che questa definizione di AWS è diversa da molte altre, in quanto considera esplicitamente le capacità di apprendimento come caratteristica fondamentale degli AWS. È questa caratteristica che orienta gran parte della discussione del capitolo.

Questo ci porta al secondo chiarimento. In base a tale definizione di AWS, gli AWS letali e non letali si distinguono sulla base dello *scopo* e non dell'effetto del loro uso. Un AWS letale è utilizzato allo scopo di esercitare una forza letale – che, cioè, avrà come risultato la morte o un trauma permanente per esseri umani. Un AWS non letale è utilizzato allo scopo di neutralizzare esseri umani "senza causare morte o traumi permanenti" (Davison, 2009, p. 1). Come si ricorderà dal capitolo 1, l'esito degli AWS è predicibile solo fino a un certo punto, perciò è concepibile che un AWS utilizzato a scopi non letali possa produrre effetti letali (Coleman, 2015; Enemark, 2008; Kaurin, 2010, 2015; Heyns, 2016a, 2016b), perché esiste "una potenziale disconnessione fra l'intenzione alla base dell'uso di un'arma e le sue conseguenze" (Enemark, 2008, p. 201). Questo scenario non invalida la distinzione che propongo qui; esemplifica, invece, proprio una delle questioni che affronto nel capitolo, cioè l'attribuzione di responsabilità morale per le azioni di sistemi imprevedibili. Parte dell'analisi della responsabilità morale degli AWS che propongo si applica allo stesso modo ai LAWS e agli AWS non letali. Distinguerò tra i due

ogniqualvolta l'analisi porti a un risultato diverso per ciascun gruppo di AWS.

Il terzo chiarimento riguarda la natura della responsabilità. Il nostro obiettivo è comprendere come un agente umano possa assumersi la responsabilità *morale significativa* per le azioni degli AWS. Questo implica che la responsabilità morale non è attribuita nominalmente agli agenti umani, per esempio per il loro ruolo o il loro grado, ma ricade in modo giustificato ed equo su chi ha avuto un ruolo fondamentale nel verificarsi degli effetti dell'uso degli AWS. Questo comporta, inoltre, che un agente umano accetti tale responsabilità morale, e la lode o il biasimo che ne conseguono, come individuo (in un senso personale) e non come membro o rappresentante di un'organizzazione della difesa o di un organismo professionale. L'attribuzione di una responsabilità morale significativa può essere alla base di processi legali per attribuire la responsabilità legale e definire chi deve rendere conto e rispondere delle azioni compiute. Anche se è collegata alla responsabilità legale, quella morale ne rimane distinta. Per esempio, come vedremo nel prossimo paragrafo, l'attribuzione della responsabilità morale, e di conseguenza la lode o il biasimo, richiedono una connessione causale e intenzionale fra un'azione e il suo effetto. Non è necessariamente così quando si considera la responsabilità legale. Si pensi, per esempio, al concetto di *faultless responsibility* (responsabilità "senza colpa") in base al quale una persona può essere punita anche se non può essere stabilita l'intenzione di commettere un reato. Questo capitolo si concentra solo sulla responsabilità morale e, anche se può preparare il terreno per l'attribuzione di responsabilità legale, non si occupa di quella.

Definito lo spazio concettuale della nostra analisi, possiamo ora affrontare la letteratura sull'attribuzione di responsabilità morale per i sistemi IA.

7.2 RESPONSABILITÀ MORALE PER I SISTEMI IA

Il dibattito sulle responsabilità morali delle azioni degli AWS è collegato alla più generale discussione sulle responsabilità morali per le azioni dei sistemi IA. Come abbiamo già detto, esiste un consenso crescente sul fatto che queste responsabilità ricadano su agenti umani; rimane problematico invece attribuirle correttamente, dato che, perché l'attribuzione di responsabilità morale sia giustificata ed equa, gli agenti devono avere un rapporto specifico con le loro azioni e le loro conseguenze. Questo rapporto deve soddisfare tutte le seguenti quattro condizioni:

- Condizione di intenzionalità: l'agente deve avere l'intenzione di ottenere un dato effetto (Kant, 2019; Branscombe et al., 1996; Khoury, 2018).
- Condizione di causalità: deve esistere una connessione causale tra la decisione/azione dell'agente e i suoi effetti (Fischer, Ravizza, 2000; Sartorio, 2007; Shoemaker, 2017).
- Condizione di consequenzialità: l'agente deve avere una comprensione degli effetti della decisione/azione, così come del loro valore morale e del biasimo o della lode che ne conseguono (Bentham, 1789; Wallace, 1998; Levy, 2008; Kelly, 2012).
- Condizione di scelta: l'agente ha qualche grado di libertà che gli consente di scegliere tra linee di condotta differenti (Strawson, 1962; Watson, 1975; Nelkin, 2011).[1]

Tutte e quattro le condizioni si rivelano problematiche da soddisfare quando si tratta di sistemi IA, e le difficoltà nel rispettarle costituiscono il fondamento del cosiddetto *responsibility gap* (Matthias, 2004; Floridi, 2012). La natura distribuita del ciclo di progettazione, sviluppo e implementazione, insieme alla limitata trasparenza che spesso caratterizza questi sistemi, può rendere arduo – se non del tutto impossibile – individuare con precisione quali azioni abbiano causato uno specifico esito. Parallelamente, l'agente umano coinvolto in una qualunque fase del ciclo di vita dell'IA può non disporre della necessaria comprensione

tecnica, degli scopi d'uso o degli effetti concreti della tecnologia per poter valutare in modo adeguato le conseguenze, secondo quanto richiesto dalla condizione di consequenzialità. La possibilità di soddisfare la condizione della scelta dipende, a sua volta, da come si definiscono i concetti di libertà e di autonomia decisionale.

Sebbene le ultime tre condizioni siano difficili da rispettare, tale difficoltà è contingente: non c'è nulla di intrinsecamente incompatibile tra l'uso dell'IA e il soddisfacimento di queste condizioni. Non è così, invece, per la condizione dell'intenzionalità, il cui adempimento si scontra con la natura intrinsecamente imprevedibile dei sistemi IA. Ed è proprio su questa condizione che concentrerò l'analisi nel resto del capitolo.

Spesso la responsabilità morale viene attribuita al fine di distribuire lode o biasimo agli individui per le loro azioni moralmente buone o cattive. Per garantire che l'assegnazione della responsabilità morale sia giustificata, un elemento fondamentale è quindi l'intenzionalità dell'agente.

> Sarebbe controproducente attribuire la responsabilità, e quindi assegnare biasimo o lode, punizioni o ricompense, se le azioni degli agenti non fossero intenzionali, perché tale attribuzione sarebbe *arbitraria* e *indistinguibile* da un'assegnazione puramente casuale, il che annullerebbe il senso di biasimo o lode, punizioni o ricompense. (Floridi, 2016, p. 4, corsivo mio)

Secondo una posizione etica classica, la mancanza di intenzionalità mina l'assegnazione di responsabilità morale, anche qualora la catena causale degli eventi che hanno portato a un dato esito fosse chiara. Il comportamento dei sistemi IA può non essere il risultato diretto delle intenzioni dei singoli progettisti, sviluppatori o utilizzatori, per due motivi: le azioni distribuite e la mancanza di predicibilità. Consideriamo il primo caso. I sistemi IA possono compiere azioni che hanno valenza morale ma derivano da molte azioni moralmente neutre – cioè, azioni individuali compiute da esseri umani o altri agenti artificiali che in sé (da sole) non portano a esiti specificamente buoni o cattivi (Floridi, 2016). Possiamo immaginare una rete di agenti coinvolti nella progettazione, nello sviluppo e nell'uso di un sistema IA, ciascuno dei quali prende decisioni neutre sul piano morale ma tali che, una volta coordinate a livello di rete, portano a un esito moralmente negativo.

Questa è quella che Floridi (2012) chiama moralità distribuita. Si può ritenere che tutta la rete sia moralmente responsabile di quelle azioni, ma

attribuire biasimo o lode a ogni singolo individuo o gruppo di agenti nella rete non sarebbe giustificato, perché le singole azioni in sé non portano ad alcun esito con una valenza morale – anche se hanno una valenza morale una volta coordinate al livello della rete. La moralità distribuita non è un fenomeno esclusivo dei sistemi IA, ma assume una rilevanza particolare proprio in questo ambito. Qui, la rete di agenti coinvolti e la frammentazione delle responsabilità lungo le diverse fasi operative sono tanto pervasive da rendere arduo individuare un'intenzionalità chiaramente attribuibile alle azioni compiute, e dunque distribuire la responsabilità morale in modo coerente.

Passando al secondo caso, come abbiamo già visto nel capitolo 1, una volta implementati, alcuni sistemi IA possono generare comportamenti nuovi, non previsti né prevedibili dai soggetti umani che li hanno progettati, sviluppati o usati. In tali circostanze, attribuire loro responsabilità morale è problematico, proprio in virtù dell'assenza di intenzionalità all'origine di quegli esiti. Per essere più precisi:

> L'intenzionalità non è chiusa rispetto all'implicazione causale […]. Nel caso diretto di non chiusura, non è detto che, se Alice intende causare *a*, e *a* causa *b*, allora Alice intenda causare *b*. (Floridi, 2016, p. 4)

Tale osservazione assume un peso particolare quando si considerano sistemi IA, i quali possono sviluppare comportamenti imprevisti e non voluti – il cosiddetto problema della prevedibilità. Questa imprevedibilità compromette il soddisfacimento della condizione delle conseguenze: se gli esiti generati dal sistema una volta implementato non possono essere anticipati, allora non è realisticamente possibile, per l'agente umano, considerare tutte le potenziali conseguenze della propria azione né valutarne il significato morale.

Come si ricorderà dal capitolo 1, sulla predicibilità influisce un insieme molto ampio di variabili: le funzionalità tecniche del sistema, le caratteristiche del contesto d'uso, il livello a cui l'operatore comprende il funzionamento del sistema e, nel campo della difesa, il comportamento degli avversari. Queste variabili possono cambiare e interagire tra loro in modi diversi, perciò diventa difficile prevedere tutte le azioni che un sistema IA potrebbe compiere e i loro effetti. La limitata predicibilità degli esiti di un sistema rende difficile, se non impossibile, collegare gli esiti alle intenzioni degli agenti umani che progettano, sviluppano e usano tali

sistemi; a sua volta, questo rende impossibile, seguendo l’approccio classico, attribuire la responsabilità morale delle azioni dei sistemi IA ad agenti umani.

7.3 RESPONSABILITÀ MORALE COLLETTIVA E DISTRIBUITA SENZA COLPA

Nel dibattito etico contemporaneo sono emerse alternative al tradizionale approccio che mira ad attribuire responsabilità morale individuale per le azioni compiute da sistemi IA. Tra queste, particolare rilievo hanno assunto le teorie della responsabilità morale collettiva (Corlett, 2001; List, Pettit, 2011) e distribuita (Floridi, 2012, 2016).

Le prospettive incentrate sulla responsabilità collettiva – come quella avanzata da A. Krishnan (2009) e, in forma più articolata, da List e Pettit (2011) – spostano l'attenzione dall'individuo al gruppo, trattando quest'ultimo come un agente unitario cui può essere attribuita responsabilità morale per le decisioni compiute in modo concertato. In un'analisi particolarmente influente, List e Pettit sostengono che:

> Un agente collettivo [cioè un gruppo di persone considerato come un'entità unica] può essere ritenuto moralmente responsabile di un'azione nella misura in cui soddisfa tre condizioni.
>
> Primo requisito. L'agente collettivo si confronta con una scelta normativa significativa, ovvero con la possibilità di compiere un'azione buona o cattiva, giusta o ingiusta.
>
> Secondo requisito. L'agente possiede la comprensione e l'accesso alle evidenze necessari per formulare giudizi normativi sulle opzioni disponibili.
>
> Terzo requisito. L'agente dispone del controllo necessario per scegliere tra le alternative. (*Ibidem*, p. 158).

Questa analisi dipende dalla premessa che il gruppo esprima la propria volontà e agisca di conseguenza. Come continuano gli autori:

> Per soddisfare la seconda condizione […] un agente gruppo deve essere in grado di formulare giudizi su affermazioni in merito al valore relativo delle opzioni che ha di fronte – altrimenti non avrebbe una comprensione normativa – e deve essere in grado di accedere alle evidenze sulle questioni collegate. (*Ibidem*)

Secondo questa impostazione teorica, il gruppo è un'entità omogenea e unitaria – si pensi, per esempio, a un collettivo di lavoratori in sciopero o a una folla che manifesta in strada – che agisce in modo coordinato e intenzionale. Tuttavia, tale modello si rivela problematico se applicato al ciclo di vita dei sistemi IA, il quale è strutturalmente distribuito e coinvolge una molteplicità di agenti eterogenei, spesso ignari del quadro

complessivo dell'azione collettiva in cui sono inseriti, e ancor meno in grado di formulare un giudizio normativo su di esso.

Attribuire un'intenzionalità al gruppo nella sua totalità rischia così di oscurare il ruolo – e le implicazioni etiche – delle azioni individuali non intenzionali che possono contribuire, in modo non trascurabile, all'operato del collettivo. Fare affidamento su tale approccio, dunque, non solo semplifica eccessivamente la complessità del fenomeno, ma può condurre a una distribuzione moralmente ingiusta delle responsabilità. Come scrive Corlett:

> Un *comportamento* collettivo è una condotta o un comportamento che è il risultato di un collettivo, anche se non delle sue intenzioni. Un'*azione* collettiva è causata dalle convinzioni e dai desideri del collettivo stesso, indipendentemente dal fatto che tali convinzioni e tali desideri possano essere attribuiti o spiegati in termini individualistici. […] Mi interessa capire se sia giustificato o no attribuire un'azione intenzionale a conglomerati formati da un numero molto più grande di elementi, come (grandi) nazioni e (grandi) aziende. Se questi conglomerati non sono agenti intenzionali, allora non possono essere soggetti a cui si possano attribuire responsabilità morali. (2001, pp. 575-576, corsivo mio)

Condivido l'analisi proposta da Corlett: attribuire intenzionalità a un gruppo, senza poterla rintracciare in ciascuno dei suoi membri, mina alle fondamenta la legittimità di ritenere il gruppo moralmente responsabile. Coloro che non condividono le intenzioni collettive finirebbero per portare il peso di una responsabilità senza giustificazione.

In netto contrasto con l'approccio collettivista, la prospettiva della moralità distribuita concentra l'attenzione sull'attribuzione di responsabilità per azioni moralmente rilevanti – siano esse buone o cattive – che emergono dalla convergenza di fattori diversi, indipendenti, moralmente neutri e privi di intenzionalità. Questa è stata definita da Floridi (2016) *responsabilità morale distribuita senza colpa*. Essa si applica a contesti nei quali è possibile ricostruire la catena causale di agenti e azioni che ha condotto a un certo esito morale, ma non è possibile attribuire a nessuno di tali agenti l'intenzione di produrre quell'esito. Di conseguenza, tutti gli agenti coinvolti vengono ritenuti moralmente responsabili, nella misura in cui hanno fatto parte della rete causale che ha determinato l'evento.

Secondo questa impostazione, per attribuire responsabilità morale non è necessario dimostrare l'intenzionalità dei singoli, bensì il fatto che

> un male si sia verificato all'interno del sistema, e che le azioni in questione lo abbiano causato; ma non è necessario stabilire se gli agenti responsabili siano stati negligenti o se abbiano effettivamente inteso causarlo. (*Ibidem*, p. 8)

L'intera rete di agenti viene così ritenuta pienamente responsabile dell'esito prodotto collettivamente. È fondamentale sottolineare che tale approccio consente la distribuzione della responsabilità morale fra gli agenti umani di una rete, ma non si propone di distribuire premi o sanzioni in relazione alle azioni del sistema. Il suo scopo principale è istituire un meccanismo di retroazione che stimoli tutti gli attori coinvolti a migliorare le prestazioni complessive del sistema: se ciascun agente è ritenuto moralmente responsabile, allora è plausibile che si comporti con maggiore cautela e consapevolezza, contribuendo così a ridurre il rischio di esiti indesiderati. Questo elemento risulta particolarmente rilevante quando si considerano gli AWS. Tuttavia, il modello di responsabilità morale distribuita senza colpa non è sufficiente per colmare in modo soddisfacente il *responsibility gap* in questo ambito. Nel caso degli AWS, l'attribuzione di colpa o merito assume un ruolo centrale e imprescindibile per mantenere la moralità della guerra.

7.4 RESPONSABILITÀ MORALE PER GLI AWS: L'APPROCCIO DELLA RESPONSABILITÀ MORALE COLLETTIVA

In letteratura sono già state proposte soluzioni per superare il *responsibility gap* per gli AWS. Alcuni approcci si basano sulla responsabilità collettiva; altri sull'idea di assegnare la responsabilità lungo la catena di comando oppure sulla responsabilità morale distribuita. In questo paragrafo esaminerò questi approcci, partendo da quello centrato sulla responsabilità collettiva.

In base all'idea della responsabilità morale collettiva, si potrebbe ritenere responsabile tutta la rete degli agenti coinvolti nel ciclo di vita dell'IA. Per esempio, sostiene Taylor: "Si può ritenere che l'organizzazione nel suo complesso abbia il controllo sull'esito e quindi sia corretto ritenerla moralmente responsabile" (2020, p. 327). Questo approccio si fonda sull'idea che la responsabilità collettiva venga attribuita a gruppi di agenti che condividono l'intenzione di compiere una data azione (List, Pettit, 2011). In effetti, si può sostenere che quanti lavorano alle fasi di progettazione, sviluppo e uso degli AWS condividono l'intenzione di sviluppare un sistema in grado di esercitare una forza (letale) in modo specifico ed entro certi vincoli, pertanto hanno la responsabilità morale delle azioni che si vuole far compiere al sistema. Per il settore privato, per esempio, questa responsabilità è simile a quella che hanno i fornitori di tecnologia per i possibili malfunzionamenti dei loro prodotti. Schulzke, per esempio, sottolinea che

> nella misura in cui le azioni degli AWS derivano da come è progettato il loro software o il loro hardware, la responsabilità delle armi autonome dovrebbe essere degli sviluppatori che le creano. Se le loro azioni sono abilitate o limitate da funzionari civili e militari nella loro catena di comando, quei funzionari devono condividere la responsabilità per le azioni delle armi autonome. (2013, p. 204)

Questo approccio è stato accettato in passato quando si trattava dello sviluppo di altre armi (Weeramantry, 1985; Glerup, Horst, 2014; Miller, 2018; Khosrow-Pour, 2021), ma non è adeguato al caso degli AWS, perché non tiene conto della possibilità che, una volta in uso, gli AWS possano sviluppare comportamenti indipendentemente dalle intenzioni dei loro progettisti, sviluppatori e utenti. Vale la pena di chiarire qui che la

responsabilità morale per questi comportamenti non voluti non è simile a quella per i malfunzionamenti dei sistemi: quest'ultima può essere riferita a negligenza, cioè è responsabilità per qualche tipo di anomalia che gli agenti umani avrebbero potuto e dovuto considerare e prevenire. Quando si considera il comportamento non voluto degli AWS, si ha a che fare con comportamenti che possono emergere anche come conseguenze *corrette* delle caratteristiche tecniche dei sistemi e/o delle loro interazioni con l'ambiente circostante, ma che non avrebbero potuto essere previste dagli agenti umani, come abbiamo visto nel capitolo 1.

Al contempo, la focalizzazione sul gruppo può portare ad attribuire responsabilità morali e biasimo/lode a organizzazioni invece che a individui. Come suggerisce Taylor:

> Si possono identificare vari gruppi distinti come potenziali *loci of responsibility*: il governo, l'esercito, gli sviluppatori di LAWS. Penso che si possano fare i maggiori passi avanti per colmare il *responsibility gap* esplorando la possibilità di attribuire la responsabilità alle organizzazioni che progettano e sviluppano i LAWS. (2020, p. 327)

Tale approccio offre una soluzione limitata, perché è problematico considerare questi gruppi come agenti *intenzionali*, come sosteneva Corlett nel passo citato sopra. Per superare questa difficoltà si può pensare di attribuire la responsabilità morale a singoli individui in quanto rappresentanti di gruppi o istituzioni (Champagne, Tonkens, 2015; Galliott, 2017). In questo caso, la responsabilità è attribuita in funzione del ruolo ricoperto – e dei doveri e obblighi che a tale ruolo sono connessi – piuttosto che sulla base delle intenzioni individuali o di un nesso fattuale diretto tra causa ed effetto (Liu, 2016). Champagne e Tonkens (2015), per esempio, propongono che figure di rango elevato, in ambito sia militare sia civile, debbano essere ritenute responsabili dell'eventuale impiego scorretto degli AWS in virtù della carica che occupano. Chi riveste un alto incarico, sostengono gli autori, "acconsente volontariamente" alle condizioni inerenti a tale ufficio e può dunque, almeno in linea di principio, essere ritenuto responsabile dell'utilizzo degli AWS, anche qualora questi si rivelino imprevedibili nel loro funzionamento.

In un certo senso, tali proposte risultano coerenti con i principi fondamentali delle democrazie liberali, in cui i privilegi dell'ufficio sono inseparabili da responsabilità precise, nonché dalla possibilità di essere ritenuti responsabili per conseguenze che non sempre sono prevedibili, né direttamente riconducibili all'agente (Haddon, 2020). Tuttavia, un'analisi

più approfondita rivela che questa forma di attribuzione, nella migliore delle ipotesi, assegna una responsabilità morale in modo meramente nominale: gli agenti umani identificati come responsabili lo sono più per via della posizione che occupano che per le intenzioni, decisioni o azioni effettivamente compiute. Questo approccio rischia di produrre capri espiatori, svuotando così la responsabilità morale del suo significato sostanziale.

Il lettore potrà forse ricordare gli atti dei processi di Norimberga, che sottolineavano l'irrinunciabile esigenza di stabilire una responsabilità individuale autentica per ogni atto illecito commesso nel contesto della guerra.

Il prossimo paragrafo analizza più approfonditamente gli approcci che si concentrano sulla distribuzione della responsabilità morale lungo la catena di comando.

7.4.1 Responsabilità morale per gli AWS: distribuire la responsabilità morale lungo la catena di comando

Gli approcci che sostengono l'idea di distribuire la responsabilità morale per le azioni degli AWS lungo la catena di comando adottano due criteri per allocare la responsabilità:

(i) in misura proporzionale al potere decisionale e all'accesso alle informazioni che caratterizzano le diverse posizioni lungo la catena di comando; e
(ii) in base alla catena di comando (all'interno di un'organizzazione militare) e quindi all'autorità che il personale di grado più elevato ha sull'autonomia del personale di grado inferiore.

L'assunto (i) è coerente con l'argomentazione proposta da Walzer, secondo la quale abbiamo standard più elevati per i comandanti non solo per la pericolosità degli strumenti che hanno a loro disposizione, ma anche perché hanno "accesso a tutte le informazioni disponibili ed anche agli strumenti che gli consentono di produrre nuove informazioni" (1977, p. 395). L'assunto (ii) segue da (i) e si rispecchia nelle regole di ingaggio che le organizzazioni della difesa definiscono prima di impegnarsi nelle operazioni. Le regole di ingaggio sono stabilite a cascata lungo i vari

livelli della gerarchia, dove ciascun livello impone un grado maggiore di vincolo – e di specificità – ai livelli che gli sono subordinati. Ciò significa che, ogni volta che si scende di un livello lungo la catena, si trova un grado sempre minore di discrezionalità e un grado maggiore di specificità delle possibili decisioni/azioni, e può essere difficile o del tutto impossibile agire in modo contrario alle decisioni prese ai livelli superiori. Il personale che occupa una posizione più elevata nella catena di comando, perciò, si assume la responsabilità per le azioni compiute in esecuzione dei suoi ordini dal personale di livello inferiore, dato che quest'ultimo non ha l'autonomia per comportarsi in modo diverso da come gli è stato ordinato. Questa idea è condivisa da molti teorici della Guerra Giusta (Walzer, 1977), anche se non da tutti (McMahan, 2006). Come scrive Walzer: "Consideriamo i soldati che eseguono degli ordini come uomini i cui atti non sono strettamente personali e la cui responsabilità per ciò che fanno risulta in qualche modo ridotta" (1977, p. 384). Questa autonomia ridotta, e perciò anche responsabilità ridotta, si rispecchia nella dottrina della responsabilità del comando, secondo la quale i superiori devono rispondere di ciò che fanno i loro subordinati, in base al principio di omissione (per esempio, per non avere prevenuto o non essere intervenuti), quando almeno in linea di principio hanno sui loro subordinati un controllo sufficiente per prevenire un comportamento immorale, o quantomeno per intervenire e limitarlo.

Per quanto riguarda gli AWS, entrambi gli assunti (i) e (ii) sono problematici, perché non tengono conto delle caratteristiche di questi sistemi e delle difficoltà pratiche e concettuali che possono derivare dal loro uso. Cominciamo con l'assunto (i): confonde l'ampiezza del potere decisionale e l'importanza delle informazioni a cui si ha accesso con il livello di granularità delle informazioni a supporto. Per esempio, in alcune circostanze i rischi e i vantaggi dell'uso di AWS in uno specifico teatro operativo possono essere più chiari al personale di grado inferiore (in particolare a quello che ha familiarità con la tecnologia e con il contesto) che al personale di grado superiore, che spesso non vaglia informazioni specifiche sulla tecnologia o sul contesto d'uso. Mentre le informazioni su rischi e benefici di un'operazione specifica possono essere trasmesse verso l'alto lungo la catena di comando, è probabile che, in un contesto in cui gli AWS sono usati in maniera routinaria, la granularità di quelle informazioni diminuisca a mano a mano che si sale di livello (Payne,

2021, pp. 110-112). Questo approccio rischia di condurre a una situazione in cui il personale ai vertici della catena di comando è ritenuto moralmente responsabile delle azioni compiute dagli AWS, pur non disponendo di informazioni sufficientemente dettagliate affinché tale responsabilità possa essere attribuita in modo giustificato ed equo. È importante sottolineare che si tratta, in primo luogo, di un problema di natura pragmatica: in linea di principio, nulla impedisce alle istituzioni militari di istituire processi adeguati per superare questo limite e favorire l'accesso alle informazioni necessarie. In tal caso, l'assunto (i) rimarrebbe valido. Tuttavia, finché non verrà implementato un sistema che garantisca ai decisori accesso tempestivo a informazioni con il grado di granularità necessario per valutare benefici e rischi dell'impiego degli AWS, l'idea secondo cui la responsabilità morale possa essere distribuita lungo la gerarchia del comando resta priva di un fondamento solido.

L'assunto (ii) pone problemi concettuali in merito a controllo e autonomia. Partiamo dal controllo. Le responsabilità morali dei comandanti per le azioni dei loro subordinati si basano su tre condizioni:

> (1) l'esistenza di una relazione superiore-subordinato, in cui il superiore ha un controllo effettivo sul subordinato; (2) l'elemento mentale necessario, che richiede in generale che il superiore sapesse o avesse motivo di sapere (o avrebbe dovuto sapere) dei crimini commessi dai subordinati; e (3) l'avere omesso di controllare, prevenire o punire i crimini commessi. (Jain, 2016, p. 310)

Stabilire un controllo sugli AWS – che sia definito "efficace", "appropriato" o "significativo" – si è rivelato un compito tutt'altro che agevole (Ekelhof, 2019). Le difficoltà riscontrate possono dipendere, in parte, dalle modalità di impiego di questi sistemi – si pensi, per esempio, alla distinzione tra controllo umano *on the loop* o *post loop* – ma anche dalle caratteristiche del contesto operativo, dalla tipologia specifica del sistema adottato, o da una combinazione di tali fattori.

Il problema della predicibilità introduce un'ulteriore complicazione: una volta attivati, gli AWS possono agire in modi imprevisti, non voluti e, in alcuni casi, persino indesiderati. Ne consegue che gli ufficiali responsabili del loro dispiegamento potrebbero trovarsi nell'impossibilità di anticipare gli esiti indesiderati delle operazioni e, dunque, di prevenirli – rendendo difficoltoso, se non impossibile, soddisfare le condizioni (1) e (2) delineate da Jain.

La condizione (3), che prevede la responsabilità dei comandanti per le azioni scorrette dei propri subordinati, si fonda sul presupposto che tali comandanti siano in grado di esercitare controllo e prevenire comportamenti impropri. Tuttavia, nel caso degli AWS, l'intrinseca imprevedibilità del loro comportamento mina alla base tale presupposto, rendendo l'attribuzione di responsabilità morale non solo problematica, ma ingiustificata.

7.4.2 Responsabilità morale per gli AWS: l'approccio della responsabilità morale senza colpa distribuita

Alcuni dei limiti degli approcci descritti fin qui si possono superare prendendo in considerazione la responsabilità morale senza colpa distribuita.

> In base a questo quadro generale, i comandanti sarebbero responsabili per le azioni degli AWS all'incirca nella stessa misura in cui lo sono oggi, poiché hanno un potere di vincolare l'autonomia degli AWS simile a quello che hanno rispetto ai soldati umani. […] L'esatta distribuzione di biasimo fra comandanti e sviluppatori può essere determinata solo nella misura in cui contribuiscono con le loro azioni (o la loro mancanza di azioni) alle azioni illegali degli AWS. (Schulzke, 2013, p. 216)

Questo approccio consente di attribuire la responsabilità morale a tutti gli individui che partecipano (determinandoli) alla progettazione, allo sviluppo e all'uso degli AWS; seguendolo, però, è difficile attribuire in modo giustificato lode o biasimo morali. I motivi sono due: mancanza di trasparenza e mancanza di intenzionalità.

Il processo di retroingegnerizzazione necessario per identificare la rete di agenti che ha definito (causalmente) il comportamento dell'AWS può essere ostacolato dalla mancanza di trasparenza del sistema stesso, o dalla limitata trasparenza e tracciabilità delle informazioni sul sistema e sul processo decisionale alla base del suo uso (Tsamados et al., 2021). Il rischio è concreto. Può nascere come conseguenza delle innumerevoli interazioni fra i molti agenti che definiscono le azioni degli AWS, che può essere difficile ricostruire con dettagli sufficienti a comprendere i fattori che hanno determinato il comportamento del sistema. Gli Stati possono anche decidere di non condividere informazioni rilevanti. Nel 2010, il giudice speciale dell'ONU sulle esecuzioni extragiudiziali, sommarie o arbitrarie, in un report sulle uccisioni mirate ha sottolineato che gli Stati

possono decidere di non usare “le salvaguardie procedurali e di altra natura esistenti per garantire che le uccisioni siano legali e giustificate, e i meccanismi di *accountability* che garantiscono che le uccisioni illecite vengano indagate, perseguite e punite” (Alston, 2010, p. 10). Come sottolineano Verdiesen, Santoni de Sio e Dignum nel loro commento al report:

> Il motivo di questa mancanza di *accountability* è che la comunità internazionale non può verificare la legalità delle uccisioni, né confermare l’autenticità dell’intelligence utilizzata nel processo di definizione degli obiettivi o garantire che le uccisioni mirate illecite non rimangano impunite. (2021, p. 145)

Allo stesso tempo, come abbiamo visto nel paragrafo 7.2, ricostruire la catena causale delle decisioni e delle azioni che hanno portato a un comportamento specifico degli AWS può non essere sufficiente per stabilire che ci sia stata l’intenzione di produrre quel comportamento. In effetti, questo approccio mira a identificare la responsabilità morale *senza colpa* e non mira ad attribuire lode o biasimo. Quindi getta luce solo in maniera limitata sul *responsibility gap* degli AWS, perché lode e biasimo sono necessari per ricompensare gli usi moralmente corretti di quei sistemi e punire, o rimediare al male morale che quegli usi possono causare.

È venuto il momento di considerare la responsabilità morale significativa per l’uso degli AWS, e la *moral gambit* (la scommessa morale) che propongo per attribuire questa responsabilità.

7.5 *MORAL GAMBIT*: LA RESPONSABILITÀ MORALE SIGNIFICATIVA E LA SCOMMESSA MORALE

Per quanto riguarda gli AWS, la responsabilità morale include la responsabilità per qualsiasi azione che porti al danno distruttivo (letale o no) che questi sistemi possono causare. Una delle condizioni perché l'uso di questi sistemi sia moralmente accettabile è che deve essere possibile attribuire in modo giustificato la responsabilità per qualsiasi danno di quel genere. Ciò è possibile quando gli esiti dei sistemi rispecchiano l'intenzionalità degli agenti umani e quando esiste un legame causale fra le decisioni/azioni degli agenti umani e gli esiti degli AWS. È importante, inoltre, che la responsabilità morale sia attribuita in modo equo (*fairly*), per esempio che gli agenti coinvolti abbiano informazioni e una comprensione del contesto in cui operano sufficienti per poter considerare tutte le possibili alternative prima di prendere qualsiasi decisione. Si può attribuire la responsabilità morale in modo giustificato ed equo solo quando sono soddisfatte tutte le quattro condizioni specificate nel paragrafo 7.2. Inoltre, gli agenti devono poter accettare questa responsabilità morale come un elemento delle loro azioni e decisioni, e devono poter ricevere la lode o il biasimo che ne derivano come una valutazione del loro carattere morale. La responsabilità morale *significativa* può essere attribuita solo se sono soddisfatti tutti questi criteri.

Attribuire la responsabilità morale significativa per le azioni degli AWS è una condizione preliminare necessaria per il loro uso, perché è il tipo di responsabilità che mostra un minimo di *due care* ("dovuta attenzione"; Strawson, 1962) per chi subisce le azioni degli AWS. In tal senso ha ragione Sparrow nel sottolineare che "il minimo che dobbiamo ai nostri nemici è riconoscere che le loro vite valgono abbastanza perché qualcuno accetti la responsabilità della loro morte" (2007, p. 67). Una responsabilità morale significativa consente di esercitare la responsabilità retrospettiva, in quanto alimenta i meccanismi di *accountability*. Al tempo stesso, può fungere da leva per una responsabilità prospettica, nella misura in cui la possibilità di ricevere lode o biasimo in seguito a una determinata scelta o

azione contribuisce a incentivare comportamenti accorti e moralmente fondati.

La limitata predicibilità degli AWS limita la responsabilità morale significativa. Sarebbe ingiustificato e non equo *attribuire* ad agenti umani la responsabilità morale per (tutte le) azioni non previste che un AWS può compiere, perché quelle azioni non derivano dall'intento dell'agente umano (la responsabilità morale non sarebbe giustificata) e nessuna informazione consentirebbe all'agente di prevedere la totalità delle azioni che un AWS potrebbe compiere una volta in uso, per identificare e prevenire quelle indesiderate (la responsabilità morale non sarebbe equa). Tutto quello che si potrebbe chiedere è che progettisti, sviluppatori e utilizzatori di AWS *si assumano* la responsabilità morale significativa per le azioni intese, essendo consapevoli del rischio che possano verificarsi esiti imprevisti, e *si assumano* la responsabilità morale per gli effetti impredicibili che possono derivare dalla decisione di utilizzare gli AWS. Chiariamo questo punto.

Nell'assumersi questa responsabilità, gli agenti umani fanno una *scommessa morale* (*moral gambit*): progettano/sviluppano/usano un AWS, pienamente consapevoli del rischio che possa compiere qualche azione imprevista. Per limitare questi rischi (e ottimizzare le probabilità di una scommessa vincente), gli agenti umani e le istituzioni della difesa rilevanti devono agire al loro meglio e stabilire tutte le possibili misure per ridurre il male morale (e favorire il bene morale) che il comportamento imprevisto può causare. Gli agenti umani rimangono consapevoli che, indipendentemente da tutti questi sforzi, non sarà possibile prevedere tutte le possibili azioni degli AWS e i loro effetti, nel contesto del loro uso.[2] Ciononostante, se decidono di procedere con progettazione/sviluppo/uso di quei sistemi, allora fanno una scommessa morale e decidono di essere moralmente responsabili degli esiti imprevisti degli AWS e dei loro effetti.

Nel caso degli AWS, tutti gli agenti umani che partecipano intenzionalmente alla progettazione, allo sviluppo e all'impiego di tali sistemi accettano quella che si potrebbe definire una scommessa morale. Il personale incaricato di autorizzarne l'impiego assume – o rifiuta – tale azzardo solo nella misura in cui possiede una reale facoltà di scelta circa l'uso o il non uso degli AWS. Ne consegue che il *responsibility gap* associato alle azioni di questi sistemi può essere colmato, a condizione

che si riesca a individuare un'intenzionalità, una relazione causale con il comportamento degli AWS, una piena consapevolezza della loro intrinseca imprevedibilità e, infine, la disponibilità ad assumersi l'onere morale di tale scelta. Come vedremo nella sezione successiva, perché ciò sia possibile è indispensabile istituire un'infrastruttura adeguata per garantire l'accesso alle informazioni rilevanti, la tracciabilità dei processi decisionali e l'esclusione di esiti letali. Prima di procedere, tuttavia, occorre precisare tre aspetti fondamentali per delineare con chiarezza i confini di questa scommessa morale.

La scommessa morale non riguarda la decisione di partecipare a un'azione i cui esiti sono relativamente imprevedibili; riguarda invece, una volta deciso di partecipare, l'*accettazione volontaria* della responsabilità morale per tutto lo spettro degli esiti possibili che possono derivare dall'uso degli AWS, che siano previsti o no. In questo senso la scommessa morale non può essere imposta; deve essere compiuta volontariamente, e la responsabilità che ne consegue viene *accettata*, non attribuita. La scommessa morale significa assumersi *ex ante* la responsabilità di qualsiasi cosa possa accadere in quella specifica operazione, sperando che non si verifichino mai esiti non intesi e non desiderati, ma dichiarandosi disposti a essere chiamati a risponderne, nel caso in cui ciò accada.

Questo ci porta al secondo punto da chiarire: l'approccio alla base della scommessa. La scommessa morale addossa agli agenti umani un onere pesante: devono accettare la scommessa e la responsabilità morale che l'accompagna. Perciò chi l'accetta deve essere pienamente informato e perfettamente consapevole dei rischi e delle conseguenze delle proprie scelte, e le istituzioni hanno il dovere fondamentale di sostenere queste persone. Questo maggior onere offre un modo per colmare il *responsibility gap*, garantendo al contempo che la responsabilità morale rimanga significativa ed equa.

Il terzo chiarimento riguarda l'ammissibilità della scommessa morale. Nel contesto della difesa nazionale, e parlando di AWS non letali, la scommessa può essere accettabile. Non lo è più quando si pensa ai LAWS. In questo caso, sarebbe una scommessa fatta sulla vita di altri, e quindi sarebbe moralmente inaccettabile.

La non accettabilità di una scommessa morale sulla vita di altri si giustifica diversamente a seconda di chi è che subisce le azioni di guerra.

Per quanto riguarda i non combattenti, la scommessa morale è immediatamente esclusa, in base al principio di distinzione. Tale principio dichiara i non combattenti immuni da attacchi in ogni fase della guerra. Dato il problema della predicibilità, non vi è certezza che i LAWS rispettino la distinzione (Blanchard, Taddeo, 2022b; vedi anche il capitolo 8) e sarebbe moralmente inaccettabile scommettere sul fatto che il sistema sia in grado di rispettare questo principio.

L'argomento che determina l'inaccettabilità della scommessa morale sulla vita dei combattenti è più complesso. Vale la pena di sottolineare che, in questo caso, il problema riguarda la modalità in cui viene tolta la vita e non l'uccisione in sé (Blanchard, Taddeo, 2022c). Si potrebbe sostenere che i combattenti rinunciano al diritto di non essere uccisi e che pertanto non fa differenza se vengono uccisi da esseri umani o da LAWS (Walzer, 1977, p. 42; Meisels, 2018, pp. 11-29) o se sia o no possibile attribuire una responsabilità morale per la loro morte. L'ammissibilità della scommessa è perciò un punto discutibile. Che i combattenti rinuncino al proprio diritto alla vita è questione oggetto di accese controversie teoriche (Kamm, 2004; McMahan, 2011; Bazargan, 2014); tuttavia, ai fini dell'argomentazione che segue, si può accettare tale ipotesi come assunta.

La questione, tuttavia, non riguarda chi possa essere legittimamente ucciso, ma con quali modalità sia moralmente accettabile colpire coloro che sono ritenuti legittimi bersagli. Anche ammettendo che i combattenti rinuncino al proprio diritto alla vita, ciò avviene sotto l'implicita aspettativa che ogni attacco nei loro confronti rispetti il principio di necessità militare e quello dell'uguaglianza morale tra combattenti. È teoricamente concepibile – per quanto altamente improbabile – che i LAWS possano essere impiegati contro combattenti umani in modo conforme al criterio della necessità. Concezione, questa, che trova una certa legittimazione nella Teoria della Guerra Giusta, secondo cui lo stato di "emergenza suprema" può giustificare l'impiego di qualsiasi mezzo qualora la posta in gioco sia la sopravvivenza stessa (Walzer, 1977, pp. 251-268). Tuttavia, questo scenario è poco plausibile: da un lato, perché simili circostanze sono estremamente rare; dall'altro, perché i vantaggi operativi attribuiti ai LAWS – come la rapidità decisionale – superano di gran lunga le capacità umane, finendo per ridurre arbitrariamente la soglia di necessità per il loro impiego contro combattenti umani.

Allo stesso tempo, l’uso di LAWS viola il principio di uguaglianza morale dei combattenti. In base a questo principio, i combattenti aderiscono a un contratto marziale che implica regole reciproche di belligeranza, per le quali la rinuncia al diritto di non essere uccisi è accompagnata da un sistema di norme. Come scrivono Skerker e colleghi:

> Si può pensare che i soldati cedano il diritto di non essere oggetto di violenza letale da parte di combattenti nemici in accordo con le norme militari che ottimizzano un compromesso fra massimizzare i propri interessi militari e minimizzare la sofferenza del nemico. (Skerker, Purves, Jenkins, 2020, p. 202)

Tra le norme a cui fanno riferimento Skerker e colleghi c’è l’aspettativa che i combattenti non vengano presi di mira in modo arbitrario o sconsiderato. Ne deriva che l’uccisione non intenzionale – sia essa di combattenti o di civili – risulta quantomeno moralmente problematica. Anche nel caso in cui i combattenti rinuncino al diritto a non essere uccisi, lo fanno nella presunzione che, se colpiti, ciò avverrà in modo deliberato, e non come esito di un comportamento imprevisto o di un azzardo morale andato storto. Gli operatori che decidono di impiegare un LAWS accettano una scommessa: confidano che il sistema agisca secondo l’uso previsto, pur essendo consapevoli che vi è la possibilità che esso identifichi, selezioni e ingaggi un bersaglio non intenzionale. Questo azzardo, sostengo, è incompatibile con le aspettative imposte dai principi di necessità e di uguaglianza morale tra combattenti. Di conseguenza, fintanto che i LAWS rimangono imprevedibili nel processo di identificazione, selezione e ingaggio di obiettivi umani, non è possibile attribuire una responsabilità morale significativa per le loro azioni – il che rende, in ultima analisi, il loro impiego moralmente inammissibile.

7.6 RESPONSABILITÀ MORALE SIGNIFICATIVA PER LE AZIONI DI AWS NON LETALI

In gran parte la discussione sugli AWS si concentra sui LAWS, perché gli usi letali degli AWS comportano perdita di vite umane e di conseguenza gravi problemi etici. Gli AWS non letali, invece, hanno attirato minore attenzione, ma credo che questa lacuna sia problematica. I rischi etici degli AWS possono essere meno gravi, per impatto e ordine di grandezza, di quelli creati dai LAWS, ma sono comunque rischi importanti. Gli AWS non letali possono provocare danni significativi, come traumi alle persone, distruzione di edifici e cose, violazioni della libertà e violazioni del principio di distinzione. Questi rischi nascono, per esempio, se gli AWS non letali vengono impiegati "quando è richiesto l'uso di una forza graduata e la forza letale è l'eccezione" (Heyns, 2016a, p. 5). Attribuire la responsabilità morale per l'uso di AWS non letali, perciò, è fondamentale. La scommessa morale qui delineata offre una proposta innovativa e quanto mai necessaria a questa impasse etica.

La scommessa morale non può essere imposta a un agente umano, né può essere conseguenza di un ruolo: perché sia accettabile, l'agente deve farla volontariamente. Nel caso dell'uso di *AWS non letali*, si possono stabilire varie procedure per sostenere la decisione di accettare, o non accettare, la scommessa morale. Qui di seguito presento otto raccomandazioni che i fornitori di tecnologia nel campo della difesa e le stesse organizzazioni della difesa dovrebbero seguire per sostenere in questo senso i loro membri. Le raccomandazioni sono elencate in ordine logico.

1. Fornire AWS i cui sistemi IA siano *interpretabili* e non semplicemente spiegabili. La non predicibilità dei sistemi IA è in parte una funzione della loro mancanza di trasparenza. Le spiegazioni di un modello a scatola nera possono offrire una rappresentazione imprecisa del modello originale (Rudin, 2019); per questo motivo la *explainability* offre soluzioni limitate ai problemi posti dalla mancanza di predicibilità. Si possono ottenere esiti migliori fornendo modelli interpretabili, cioè un modello "vincolato nella forma di modello così che o sia utile a qualcuno, oppure incorpori la

conoscenza strutturale del dominio, come monotonicità, causalità, vincoli strutturali (generativi), additività o vincoli fisici che derivano dalla conoscenza del dominio" (*ibidem*, p. 206).

2. Valutare la predicibilità. I fornitori e le istituzioni della difesa devono esaminare attentamente quali caratteristiche tecniche e operative dei sistemi d'arma autonomi incidano negativamente sulla loro capacità di identificare, selezionare e ingaggiare obiettivi in modo prevedibile, e impegnarsi a migliorarle al fine di ridurre al minimo esiti inattesi.

3. Coloro che sono chiamati a decidere sull'impiego di sistemi d'arma autonomi non letali dovrebbero possedere un elevato grado di competenza tecnica, sia in merito al funzionamento di tali sistemi, sia rispetto al contesto operativo in cui verranno impiegati. Solo una comprensione approfondita consente di coglierne appieno le potenzialità strategiche e tattiche, di valutarne l'uso ottimale, nonché di anticiparne eventuali vulnerabilità. Questa preparazione tecnica costituisce la base imprescindibile per una scelta informata: solo così è possibile ponderare adeguatamente i rischi connessi a comportamenti imprevedibili e le implicazioni morali che l'assunzione della scommessa etica inevitabilmente comporta.

4. Tracciabilità dei processi. Le informazioni sulle specifiche tecniche di sistemi d'arma autonomi non letali, del loro ciclo di sviluppo e della loro modalità di fornitura devono essere trasparenti ai decisori. Nello stesso modo, qualsiasi informazione rilevante in merito al sistema che possa migliorare la comprensione che ne hanno i decisori deve essere trasmessa al personale tempestivamente e in modo accurato.

5. Giustificazione degli usi. La decisione di usare o no AWS non letali deve sempre essere il frutto di un'attenta analisi costi-benefici e trovare giustificazione nel rispetto del principio di necessità. A ciò va aggiunta una valutazione ulteriore, non meno cruciale: la sicurezza del personale militare operante nel teatro specifico di riferimento.

6. Garantire effetti non letali. Devono essere predisposte misure adeguate volte a ridurre al minimo i rischi di esiti letali derivanti dall'impiego di sistemi d'arma autonomi non letali. Tali misure, pur nella loro specificità, potrebbero non discostarsi significativamente da quelle previste per l'uso

delle armi convenzionali, e potrebbero includere, tra le altre, valutazioni di necessità e proporzionalità, un'analisi attenta del contesto operativo, nonché soluzioni ingegneristiche relative all'interfaccia d'uso – per esempio, la possibilità di disattivazione remota del sistema.

7. Riparare e rimediare. È necessario istituire un processo strutturato per l'identificazione degli errori e degli esiti indesiderati, volto a valutarne l'impatto e i costi, nonché a definire misure appropriate di riparazione e rimedio. Tali misure non assolvono né annullano l'attribuzione di lode o biasimo morale nei confronti degli agenti umani coinvolti, ma costituiscono un canale attraverso cui le istituzioni della difesa possono esercitare in modo responsabile la propria rendicontazione rispetto alle decisioni assunte dal proprio personale.

8. Auditing. Deve essere introdotta una forma di audit etico, tanto sui sistemi d'arma autonomi non letali quanto sui processi che ne regolano acquisizione e impiego (Mökander, Floridi, 2021). Lo scopo è duplice: da un lato rafforzare i meccanismi di *accountability*, dall'altro individuare tempestivamente eventuali vulnerabilità operative o decisionali, così da correggerle e migliorare sia la qualità del processo decisionale, sia l'efficacia delle misure di rimedio.

7.7 CONCLUSIONE

In questo capitolo ho sostenuto che, affinché l'impiego degli AWS possa dirsi eticamente accettabile, è essenziale attribuire una responsabilità morale significativa agli agenti umani coinvolti. Tale attribuzione si basa sul soddisfacimento di requisiti stringenti, giustificati dalla gravità dei danni – letali o no – che questi sistemi possono arrecare.

Come abbiamo visto nelle pagine precedenti, tali condizioni non possono essere soddisfatte nel caso dei LAWS; di conseguenza, il loro impiego si configura come moralmente inaccettabile. Al contrario, una responsabilità morale significativa può essere concepita nel caso di AWS non letali, attraverso la strategia morale della *gambit*, a condizione che le istituzioni della difesa mettano in atto i necessari meccanismi di supporto. Tali meccanismi devono assistere gli agenti umani disposti ad assumersi la *moral gambit* nel comprendere a fondo il funzionamento degli AWS, i benefici e i rischi connessi al loro impiego, nonché le implicazioni etiche di tale assunzione di responsabilità.

Spero che le otto raccomandazioni offerte in questo capitolo possano fornire un contributo concreto a quanti, tra i fornitori di tecnologia e le organizzazioni della difesa, si impegnano a operare in questa direzione.

---

1. Nel resto di questo capitolo non prenderò in considerazione la condizione di scelta, perché nella letteratura rilevante si riferisce alle idee metafisiche di determinismo e libertà. La condizione riguarda il fatto che gli esseri umani siano pienamente determinati oppure no e, quindi, possano o no scegliere fra linee di condotta alternative. La risposta alla domanda, se un sistema IA soddisfi questa condizione indipendentemente dalle sue caratteristiche e dal suo livello di perfezionamento, è più legata alla concezione metafisica che ciascuno sostiene.

2. Come già detto nel capitolo 1, l'impredicibilità è vincolata entro confini predeterminati come payload o raggio d'azione.

8

LA TEORIA DELLA GUERRA GIUSTA E L'AMMISSIBILITÀ DEI SISTEMI D'ARMA AUTONOMI

8.1 INTRODUZIONE

È venuto il momento di prendere in considerazione il tema dell'ammissibilità morale degli AWS. Le discussioni sull'ammissibilità, sul piano morale e giuridico, degli AWS iniziano nel 2012, con la pubblicazione della direttiva del DoD degli Stati Uniti sull'autonomia nei sistemi d'arma (US Department of Defense, 2012), di cui è stata pubblicata una versione aggiornata nel 2023. Negli undici anni trascorsi fra le due versioni, quella che era una discussione speculativa su possibili, ma futuribili, usi di questi sistemi d'arma, si è trasformata in un dibattito urgente per fissare le condizioni per l'uso legittimo di queste armi, che hanno ormai fatto il loro ingresso nei campi di battaglia. Nel marzo 2021 l'ONU ha pubblicato un documento in cui si dava notizia del primo uso ufficiale di AWS sul fronte libico (Choudhury et al., 2021). Il rapporto sottolinea che quei sistemi "erano programmati per attaccare bersagli senza richiedere una connettività di dati fra l'operatore e la munizione; in effetti, una vera capacità *fire, forget and find*" (*ibidem*, p. 17). Se l'uso degli AWS sul fronte libico può essere considerato il primo (e forse isolato) caso, l'ampio schieramento di queste armi da entrambe le parti durante la guerra in Ucraina (US Department of Defense, 2022a; Tiwari, 2023; Knight, 2022) ha rotto il tabù sull'uso degli AWS e ha reso ancora più urgente la necessità di definire regolamentazioni per questo tipo di armi.

In tale contesto è problematico che il dibattito sull'ammissibilità dell'uso di AWS non abbia fatto passi avanti con la stessa rapidità ed efficienza con cui è proceduto lo sviluppo di queste armi, e pertanto che i primi usi noti di queste armi in Libia e in Ucraina siano avvenuti in un vuoto regolamentativo. Questo vuoto è il risultato di vari fattori, fra i quali la polarizzazione del dibattito sull'ammissibilità degli AWS. Dato che

l’adozione degli AWS si va diffondendo, è imperativo uscire dallo stallo e trovare un insieme condiviso di assunti normativi e di esiti desiderabili che possa guidarci nella comprensione dei casi limite dell’uso degli AWS – quegli usi, cioè, i cui esiti non sono accettabili sul piano morale e perciò devono essere proibiti – e nella definizione di norme per gli altri casi che possono essere considerati accettabili sul piano etico e giuridico. Vale la pena di precisare che sostenere la necessità di un dibattito più equilibrato sugli AWS non vuol dire rassegnarsi all’irrilevanza della questione della loro ammissibilità, dal momento che questi sistemi sono *già* impiegati sul campo. Al contrario, il punto è che, data l’adozione crescente – attuale e prevedibile – degli AWS, la questione della loro liceità deve essere affrontata e risolta sulla base di un consenso più ampio, tale da permettere l’applicazione concreta di qualsiasi misura ne derivi, sia essa un divieto totale o una regolamentazione d’uso. In quest’ottica, ogni risposta a tale quesito deve muovere da premesse che possano essere condivise da entrambe le parti del dibattito.

La Teoria della Guerra Giusta è centrale per questo dibattito e per la sua polarizzazione: alcuni sostengono che gli AWS possono essere utilizzati rispettandone i principi e le conseguenti norme dell’IHL (Arkin, 2009; Anderson, Reisner, Waxman, 2014); mentre altri sostengono che il loro uso viola quegli stessi principi e i valori fondamentali delle nostre società e pertanto non deve essere permesso (Marchant et al., 2011; Grut, 2013; Foy, 2014; Roff, 2015; van den Boogaard, 2015; Beard, 2018; Davison, 2018; Winter, 2018, 2020). Questa polarizzazione deriva da interpretazioni radicalmente diverse della Teoria della Guerra Giusta: una privilegia il suo aspetto consequenzialista, l’altra si concentra invece su quello deontologico.

Il modo in cui chi è a favore e chi è contro gli AWS interpreta il principio di distinzione è un esempio paradigmatico delle differenze nell’interpretazione della Teoria della Guerra Giusta e delle conseguenze che ne derivano. Il principio di distinzione stabilisce la protezione dei non combattenti in guerra e impone che, nel pianificare ed eseguire operazioni belliche, questi non devono mai costituire un bersaglio. Qualcuno sostiene che gli AWS possono rispettare il principio di distinzione meglio degli agenti umani e, a sostegno di questa tesi, spesso sono presentate due motivazioni. La prima è che gli AWS non hanno un senso di autoconservazione, perciò possono agire in modo più prudente quando

l’identificazione del bersaglio è incerta. Se nel sistema è codificato il principio *primum non nocere*, come scrive Arkin (2018, p. 319), gli AWS si esporrebbero maggiormente a rischi per sé stessi, pur di proteggere i non combattenti. La seconda motivazione è che gli AWS possono ridurre i casi di errata identificazione, perché, a differenza degli esseri umani, non cercano di adattare o distorcere le informazioni per farli rientrare in schemi familiari (Arkin, 2009).

Questa posizione si basa su una lettura semplicistica del ragionamento etico e dell’elemento consequenzialista della Teoria della Guerra Giusta: riduce il primo a un semplice insieme di regole, escludendo qualsiasi processo riflessivo che includa consapevolezza e riconoscimento del contesto e degli altri agenti coinvolti; esclude poi dal secondo qualsiasi riflessione sull’impatto che le azioni possono avere su chi ne è fatto oggetto. Questo approccio semplicistico si accompagna a una valutazione ottimistica delle capacità, della predicibilità e della robustezza dei sistemi IA – un ottimismo che supera di gran lunga lo stato attuale di sviluppo di questa tecnologia. Per esempio, Arkin sostiene che, se i principi della Teoria della Guerra Giusta sono codificati correttamente, gli AWS saranno in grado di rispettarli. Come abbiamo visto nel capitolo 1, i sistemi IA rimangono vulnerabili agli attacchi e la loro robustezza e predicibilità sono limitate, il che mina l’idea che si comportino necessariamente secondo le intenzioni di progettisti, sviluppatori e utenti, per esempio che rispettino sempre i principi della Teoria della Guerra Giusta. La combinazione di una concezione semplicistica e di una valutazione ottimistica banalizza i problemi etici e tecnici, elimina ogni sfumatura e trascura i possibili compromessi e i rischi legati all’uso degli AWS. Così facendo, contribuisce a una polarizzazione del dibattito sulla loro ammissibilità.

Una posizione altrettanto polarizzata deriva dalle analisi che fanno leva sull’elemento deontologico della Teoria della Guerra Giusta per opporsi all’uso degli AWS. In base a questa posizione, la Teoria della Guerra Giusta non è pienamente consequenzialista (Moseley, 2011) e ha al centro il dovere di riconoscere e rispettare la dignità degli esseri umani. Il rispetto, ossia il riconoscimento del valore unico dell’altro sul campo di battaglia, diventa quindi un elemento determinante per la moralità della guerra, che gli AWS non possono soddisfare perché riconoscere la dignità di un essere umano e il suo valore intrinseco è diverso dall’identificare un bersaglio.

Per esempio, Asaro sostiene che, perché un'uccisione sia legittima, i combattenti devono comprendere la reciproca decisione di entrare in guerra e rispettarla riconoscendo il valore significativo dell'autodeterminazione dell'avversario, inoltre devono essere in grado di riflettere sulle ragioni che giustificano le loro uccisioni e devono condividere quelle ragioni (Asaro, 2020). In modo analogo, Sparrow sviluppa la concezione di Nagel (1972) e sostiene la necessità di una relazione diretta fra combattente e bersaglio (Sparrow, 2016), perché una relazione di quel tipo consente di apprezzare e riconoscere un potenziale bersaglio come un individuo autonomo, il che è essenziale per rispettare il principio di distinzione. Secondo Sparrow, gli AWS violano il principio di distinzione perché spezzano quella relazione:

> Quanto più i sostenitori delle armi robotiche ne esaltano la capacità di prendere decisioni complesse senza input forniti da un operatore umano, tanto più risulta difficile credere che gli AWS colleghino chi uccide e chi è ucciso in misura abbastanza diretta da sostenere la relazione interpersonale che, come sostiene Nagel, è essenziale per il principio di distinzione. (*Ibidem*, p. 108)

Concordo con questa posizione. Tuttavia, anch'essa conduce a una polarizzazione del dibattito, perché non affronta le ragioni legittime avanzate a sostegno dell'uso degli AWS: superiorità militare, valore di deterrenza, efficacia, sicurezza dei combattenti e possibilità di concludere le guerre più rapidamente e pertanto di ridurre i rischi per i non combattenti. La polarizzazione del dibattito favorisce l'assenza di regolamentazione.

Per sviluppare una regolamentazione efficace è importante stabilire un insieme condiviso di assunti, valori e principi che consentano un dibattito costruttivo su quali usi degli AWS siano accettabili sul piano morale (ammesso che ce ne siano). Identificare quel terreno comune è l'obiettivo di questo capitolo. In quanto teoria etica specifica del dominio bellico, la Teoria della Guerra Giusta può offrire il fondamento condiviso necessario; la questione dirimente, tuttavia, è come interpretarla per evitare la polarizzazione che ho appena descritto.

In questo capitolo propongo un'interpretazione della Teoria della Guerra Giusta che cerca un equilibrio fra i suoi elementi consequenzialisti e il rispetto dovuto agli esseri umani in quanto tali, con la loro intrinseca dignità. A tal fine mi baserò sull'analisi che Blanchard e io abbiamo sviluppato in altre sedi riguardo alla Teoria della Guerra Giusta e agli AWS

(Blanchard, Taddeo, 2022a, 2022b, 2022c). Comincerò esaminando i principi di ultima istanza e di proporzionalità dello *jus ad bellum*, per mostrare che, se applicati al caso degli AWS,[1] tali principi non offrono indicazioni sufficientemente rilevanti per dirimere la questione della loro ammissibilità morale.

Mi concentrerò poi sul principio di necessità dello *jus in bello*, sostenendo che esso è stato colpevolmente trascurato nella letteratura di riferimento e che, al contrario, riveste un ruolo cruciale nell'identificazione degli usi moralmente consentiti di queste tecnologie. In seguito, analizzerò il principio di distinzione, per argomentare che un'interpretazione capace di bilanciare la necessità militare con gli elementi deontologici e consequenzialisti della Teoria della Guerra Giusta sia non solo possibile, ma necessaria. Questo equilibrio, suggerisco, consente di valutare i casi limite in cui l'impiego degli AWS può risultare giustificabile, offrendo al contempo criteri per una regolamentazione responsabile e circostanziata degli altri usi che si possano considerare moralmente ammissibili.

8.2 *JUS AD BELLUM* E AWS

Entrambe le parti nel dibattito sull'ammissibilità degli AWS fanno riferimento allo *jus ad bellum* (la parte della Teoria della Guerra Giusta che definisce i principi e le condizioni in cui uno Stato può fare ricorso alla guerra o all'uso della forza), in particolare ai principi di ultima istanza e di proporzionalità. Entrambe si basano su interpretazioni di senso comune di questi principi. Come ho illustrato altrove (Blanchard, Taddeo, 2022c), gli argomenti fondati sullo *jus ad bellum* non offrono basi sufficientemente solide per giustificare né l'impiego né il divieto degli AWS. L'obiettivo di questo paragrafo è sgomberare il campo da tali tesi, così da orientare il dibattito verso aspetti più pertinenti della Teoria della Guerra Giusta.

Partiamo dal principio di ultima istanza, secondo il quale la guerra deve essere l'ultima opzione da prendere in considerazione.[2] Il principio prevede che, per affrontare una data minaccia, i leader politici valutino tutti i mezzi diversi dalla guerra a loro disposizione, e che optino per strumenti non violenti (Coverdale, 2004, pp. 258-259). Vale la pena di ricordare che il principio di ultima istanza non significa che i leader politici debbano ricorrere alla guerra quando non è disponibile alcuna altra linea d'azione. Interpretare l'ultima istanza in questo modo significherebbe che la guerra non può mai essere giustificata, perché è sempre possibile dire che non tutte le alternative sono state tentate. Il principio invece "richiede un giudizio ponderato sulla possibilità che qualche alternativa immaginata abbia buone probabilità di evitare la guerra. Non richiede che ogni idea debba essere effettivamente perseguita sino alla fine" (Allen, citato *ibidem*, p. 259). Quanti obiettano all'uso degli AWS sostengono che questi ridurrebbero i costi economici, politici e umani della guerra, rendendo più conveniente quest'opzione e incentivando i leader politici a dichiarare guerra invece di sondare vie alternative per risolvere un conflitto come prescritto dal principio di ultima istanza (Asaro, 2008).

Tuttavia, il principio di ultima istanza impone che il decisore politico consideri innanzitutto mezzi alternativi, disponibili ed efficaci. Pertanto, indipendentemente dal fatto che gli AWS possano incentivare il ricorso al

conflitto, se tali alternative non sono state adeguatamente esplorate, la decisione di entrare in guerra resta ingiustificata. Ne consegue che l'obiezione secondo cui gli AWS violerebbero il principio di ultima istanza riguarda meno la natura di tali sistemi e molto di più la volontà – o la sua mancanza – da parte dei leader politici di attenersi a questo principio. L'analisi del rapporto fra tecnologia e Teoria della Guerra Giusta dovrebbe interrogarsi non solo sui dilemmi etici sollevati dall'innovazione tecnologica, ma anche su come quest'ultima incida sulle valutazioni relative alla proporzionalità e all'ultima istanza. In tal senso l'obiezione sopra menzionata agli AWS conserva una certa rilevanza nella misura in cui le nuove evoluzioni tecnologiche richiedono una vigilanza attenta per cogliere effetti imprevisti e indesiderati, come il rischio di escalation (Allenby, 2013).

Spesso entrambe le parti nel dibattito sugli AWS citano la precisione come altro elemento che influisce sulla valutazione *ad bellum*, in particolare in merito alla proporzionalità. Tutti fanno riferimento alla proporzionalità *ad bellum* e all'impatto degli AWS sulla probabilità che i leader politici che possono utilizzare gli AWS decidano di entrare in guerra. Da un lato, si pensa che gli AWS riducano la durata delle ostilità e quindi la potenziale entità dei danni inflitti ai non combattenti – e di conseguenza favorirebbero il rispetto della proporzionalità *ad bellum*. Dall'altro, chi è contrario all'uso degli AWS concorda sulla precisione di questi sistemi come fattore che riduce i danni causati dalla guerra, ma sostiene che proprio per questo abbassa le barriere al conflitto armato (Enemark, 2011; Brunstetter, Braun, 2013; Asaro, 2008; Roff, 2015), il che può condurre a un aumento nel numero delle guerre (Abney, 2013, p. 340).

In entrambi i casi, le analisi si fondano su una comprensione superficiale della proporzionalità *ad bellum*, perché confondono la proporzionalità, una valutazione contestuale, con la precisione, una proprietà oggettiva delle armi (Braun, Brunstetter, 2013).[3] La proporzionalità *ad bellum* stabilisce che, quando i leader politici decidono se entrare in guerra, i danni conseguenti alla decisione debbano essere proporzionati al bene che ci si aspetta come risultato (Hurka, 2005, p. 35). Richiede quindi una prospettiva globale, e le stime devono includere le previsioni di morti e danni strutturali ed economici. La precisione di un'arma è fondamentale nella valutazione di proporzionalità *in bello*, ma è irrilevante in quella *ad bellum*, che è una valutazione degli scopi

complessivi perseguiti con il conflitto armato, delle tattiche, dei tipi di costi previsti e così via.

Questa valutazione è già difficile e abbastanza indeterminata quando si considerano guerre prevedibili; la difficoltà non fa che aumentare quando si devono includere gli effetti dell'uso di una tecnologia specifica. Come scrivono Sechser e colleghi: "Estrapolare dalle tendenze tecnologiche correnti è problematico, sia perché spesso le tecnologie non sono all'altezza delle promesse, sia perché spesso hanno effetti di contrapposizione o di condizionamento che possono temperarne le conseguenze negative" (Sechser, Narang, Talmadge, 2019, p. 728). I tentativi di effettuare questa valutazione per guerre che ancora non sono avvenute e non sono prevedibili sono speculativi, e per questo non danno sostegno ad argomentazioni sulla probabilità di incidenza della guerra in caso di disponibilità degli AWS.

Per queste ragioni lo *jus ad bellum* non offre uno strumento valido per dirimere questioni relative all'ammissibilità degli AWS. Gli argomenti a favore o contro l'impiego degli AWS che si basano su questa parte della Teoria della Guerra Giusta riguardano più le attitudini politiche dei decisori che la portata normativa dei suoi principi. Di conseguenza, illuminano con efficacia le dinamiche decisionali nell'arena internazionale, ma non gettano alcuna luce sulla legittimità morale degli AWS in quanto tali. Il modo in cui i leader politici giungono alle proprie decisioni in contesto bellico – e l'influenza che su di essi hanno la tecnologia e le narrazioni che l'accompagnano – costituisce indubbiamente un tema di grande rilevanza analitica, ma non attiene direttamente alla questione della liceità morale degli AWS. A tal fine è sul terreno più solido dei principi dello *jus in bello* che occorre spostare l'attenzione. A essi dedicherò i prossimi due paragrafi.

8.3 *JUS IN BELLO*: IL PRINCIPIO DI NECESSITÀ

I principi di necessità, distinzione e proporzionalità sono centrali nello *jus in bello*, la parte della Teoria della Guerra Giusta che fissa i principi per la condotta delle ostilità in guerra. Il principio di necessità consente misure che siano necessarie per raggiungere un obiettivo militare legittimo e implica al contempo una concessione e un vincolo: consente l'uso della forza, ma solo entro i limiti dettati dalla stretta necessità operativa. La concessione prevista dal principio stabilisce che le misure adottate sono giustificate se – e solo se – risultano al tempo stesso lecite, ossia conformi ai principi di distinzione e proporzionalità, e strettamente necessarie al raggiungimento dell'obiettivo militare prefissato. Nella formulazione di Lackey, "il principio [di necessità] non dice che tutto ciò che è necessario è ammissibile, ma che tutto ciò che è ammissibile deve essere necessario" (Lackey, citato in Ohlin, May, 2016, p. 77). Il vincolo stabilisce che misure non necessarie non sono giustificate e devono essere evitate (Matsumoto, 2020). Il vincolo, perciò, introduce il requisito della forza minima, che impone ai combattenti l'obbligo morale di usare la forza minima necessaria per raggiungere un obiettivo militare legittimo (McMahan, McKim, 1993, pp. 516-517; Lango, 2010, p. 482).

Nel caso degli AWS, il principio di necessità trova una corrispondenza diretta nel principio di uso giustificato dell'IA che ho presentato nel capitolo 1, che prescrive di evitare sia l'abuso – che può generare nuovi rischi – sia un impiego eccessivamente cauto o limitato, che comporterebbe costi in termini di opportunità mancate (Taddeo et al., 2021).

In questo senso, il principio di necessità offre una guida essenziale per valutare l'ammissibilità degli AWS: tiene conto dei rischi che i combattenti subiscano danni e li bilancia con la necessità di raggiungere obiettivi militari. Ciononostante, nella letteratura in materia, il principio di necessità è stato trascurato (Grut, 2013; Wagner, 2014), e alcuni sostengono che non si applichi agli AWS o che non chiarisca in modo significativo l'ammissibilità di questi sistemi d'arma. Per esempio, Foy considera il principio di necessità privo di utilità per valutare l'uso ammissibile degli AWS, sostenendo che "mentre l'uso degli AWS chiama in

causa i principi della necessità militare e della sofferenza non necessaria, questi principi sono chiamati in causa in modo diverso rispetto ai principi di distinzione e proporzionalità" (2014, p. 54). Questo perché, secondo l'analisi di Foy, l'autonomia dei sistemi d'arma non incide sostanzialmente sul principio di necessità. L'obiezione però è mal posta, perché il principio di necessità e il requisito della forza minima comportano qualche livello di controllo sugli effetti di un'arma, che sia autonoma o no (Blanchard, Taddeo, 2022a). Per questa ragione devono essere applicati agli AWS non meno che ad altri mezzi e metodi bellici. Al contempo, il controllo degli effetti è un aspetto centrale per l'ammissibilità degli AWS (Tsamados, Taddeo, 2023) e tenerne conto entro i vincoli normativi del principio di necessità permette di capire quali livelli di controllo siano necessari perché gli AWS siano ammissibili.

Per comprendere il requisito della forza minima è importante considerare come il principio di necessità si differenzia dai principi di proporzionalità e distinzione e come interagisce con questi. Spesso viene confuso con il principio di proporzionalità (Chengeta, 2016, p. 131), ma i due principi impongono calcoli diversi. Un atto di guerra può essere proporzionato (perché i suoi costi sono tollerabili rispetto ai suoi benefici) e non essere necessario, perché quei benefici si sarebbero potuti ottenere con mezzi meno costosi (Hurka, 2008, p. 128). Viceversa, un atto di guerra può essere necessario se è il mezzo meno distruttivo (o magari l'unico mezzo) per raggiungere un obiettivo dato, ma non essere proporzionato, perché i costi non sono tollerabili, rispetto ai benefici.

I principi di necessità e proporzionalità riguardano attori diversi. Il secondo si riferisce ai danni non voluti ma prevedibili inflitti ai non combattenti, mentre il primo considera i danni inflitti ai combattenti (McMahan, 2009, pp. 19-23). La proporzionalità, quindi, non rende ridondante la necessità; per questo le obiezioni che ritengono ridondante il principio di necessità nel valutare l'ammissibilità degli AWS sono errate. Per esempio, Schmitt e Thurnher ritengono che sia possibile definire gli usi ammissibili degli AWS sulla base dei soli principi di proporzionalità e distinzione, senza bisogno di considerare quello di necessità:

> Per quanto riguarda le proibizioni basate sull'uso, il requisito che gli obiettivi militari portino qualche vantaggio militare renderebbe ridondante qualsiasi condizione distinta per la necessità militare. Per quanto riguarda le situazioni che sollevano problemi di proporzionalità, qualsiasi attacco che non porti un vantaggio militare ma causi danni ai civili o a oggetti civili violerebbe la regola […] la legge del conflitto armato già proibisce

> gli attacchi a quanti si sono arresi o sono comunque fuori combattimento. Se si considerano insieme queste osservazioni, il risultato è che la necessità militare ha poca o nessuna valenza indipendente, quando si valuta la legalità dei sistemi d'arma autonomi e il loro uso. (2012, pp. 258-259)

L'assunto alla base di questa obiezione è sbagliato, perché distinzione, proporzionalità e necessità riguardano rischi diversi. I tre principi influiscono l'uno sull'altro e offrono una guida sufficiente per i combattenti solo se vengono considerati insieme. La necessità è di rilevanza particolare per quanto riguarda gli AWS, perché, concentrandosi sui danni inflitti ai combattenti, affronta questioni legate al rispetto e ai rischi dei combattenti, che sono centrali per il dibattito sugli AWS e non sono affrontate dai principi di proporzionalità e distinzione.

La questione in esame riguarda gli obblighi morali che combattenti, sebbene nemici, hanno gli uni nei confronti degli altri. Se si esclude il principio di necessità, ne deriva che un combattente nemico (che non si arrende né è fuori combattimento, *hors de combat*) potrebbe essere soggetto a qualsiasi livello di violenza. Questo è precisamente ciò che il requisito di forza minima intende prevenire, poiché stabilisce un obbligo morale per i belligeranti di impiegare esclusivamente la forza strettamente necessaria per raggiungere un obiettivo militare legittimo (McMahan, McKim, 1993; Lango, 2010). Il requisito è definito nel preambolo della Dichiarazione di San Pietroburgo del 1868, dove si considera:

> – che i progressi della civiltà devono produrre l'effetto di attenuare, nei limiti del possibile, le calamità della guerra;
> – che il solo scopo legittimo che gli Stati devono prefiggersi durante la guerra è indebolire le forze militari del nemico;
> – che a tal fine è sufficiente *mettere fuori combattimento* il più gran numero possibile di nemici;
> – *che si va al di là dello scopo anzidetto se si usano armi che aggravano inutilmente le sofferenze degli uomini messi fuori combattimento o ne rendono la morte inevitabile*;
> – che l'uso di tali armi sarebbe pertanto contrario alle leggi dell'umanità. (International Committee of the Red Cross, 2020, corsivo mio)[4]

Tutto ciò che va al di là del minimo richiesto non è ammissibile. Il requisito riguarda gli usi della forza sia letali sia non letali e comporta un obbligo morale a usare mezzi non letali per neutralizzare un combattente nemico quando mezzi letali non sono necessari – in vista di un obiettivo militare legittimo (Kaurin, 2010). Per questo Childress sostiene che

l’obiettivo primario di un belligerante “non è uccidere e nemmeno ferire qualsiasi persona particolare, ma neutralizzarla o fermarla” (citato in Lango, 2010, p. 484). Con il requisito della forza minima, la Teoria della Guerra Giusta definisce un requisito per mezzi non letali. Perciò non tenerne conto in relazione all’ammissibilità morale degli AWS significa ignorare l’interrogativo se sia possibile utilizzare gli AWS a una soglia al di sotto della letalità.

8.3.1 Il principio di necessità e gli AWS

Può essere difficile distinguere tra usi letali e usi non letali di un’arma. Per citare un rapporto del 1972 sulle armi non letali della National Science Foundation degli Stati Uniti:

> Tutte le armi […] creano qualche rischio primario o secondario di morte o di traumi permanenti. La probabile gravità dei loro effetti (la loro letalità) dipende da molti fattori, non tutti determinati dalla loro progettazione. Le armi non pensate per uccidere o infliggere traumi permanenti, se usate con un certo grado di regolarità, senza dubbio causerebbero qualche decesso, per le differenze fisiologiche che esistono fra coloro nei cui confronti sono utilizzate, per malfunzionamenti fisici, per qualche impiego improprio e per altre circostanze. (Citato in Davison, 2009, p. 1)

La complessità degli ambienti in cui si combatte peggiora l’indeterminatezza degli effetti. La distinzione fra usi letali e non letali, perciò, si basa sulla valutazione del loro inteso scopo d’uso (Ramsey, 2002).

Per rispettare il principio di necessità e soddisfare il requisito della forza minima, l’intenzione del combattente deve essere supportata da giudizi legati al contesto sull’applicazione del principio di necessità nel contesto specifico (Ohlin, May, 2016, p. 86). Qui il punto è che la necessità militare in parte è governata dal principio del successo in un dato insieme di circostanze: se un’azione non aumenta la probabilità di raggiungere un obiettivo militare, allora quell’azione non è necessaria per raggiungerlo. Il successo ha due componenti: fortuna e abilità. La fortuna dipende dal fatto che le circostanze al di fuori del controllo dei combattenti umani siano favorevoli al raggiungimento dell’obiettivo. L’abilità è la capacità del combattente di raggiungere un obiettivo tenendo conto delle circostanze; richiede perciò la consapevolezza della propria capacità di raggiungere quell’obiettivo, e la capacità di tenerne conto nei calcoli della possibilità di giungere a buon fine (Ohlin, May, 2016). Come

si è detto nel paragrafo precedente, per quanto riguarda gli AWS emergono difficoltà quando si intraprende una simile valutazione, a causa della limitata predicibilità di quei sistemi.

Il problema della predicibilità in tale contesto ha due implicazioni importanti. La prima riguarda l'intenzione del combattente umano di usare AWS nell'applicazione della forza, letale o non letale. Come abbiamo visto nel capitolo 7, la limitata predicibilità del sistema autonomo separa l'intenzione del combattente dall'esito dell'uso di AWS. Un combattente può utilizzare uno di questi sistemi non avendo l'intenzione di esercitare una forza letale, ma l'AWS può comportarsi in modo tale da produrre effetti letali. Ciò mina la possibilità di utilizzare questi sistemi nel rispetto del requisito di forza minima. La seconda implicazione è relativa all'abilità del combattente di determinare le probabilità di successo con l'uso di AWS in contesti specifici. Il successo dipende dall'abilità e questa richiede una valutazione della propria capacità di raggiungere un obiettivo militare e una capacità di tenerne conto nei calcoli delle probabilità di successo. Questa valutazione però è minata dall'incertezza degli esiti dell'uso di AWS in qualsiasi contesto dato. Finché non si affrontano tali incognite, sarà problematico utilizzare gli AWS nel rispetto del requisito di forza minima, in particolare in circostanze in cui quel requisito richiede l'uso di una forza non letale, che impone un controllo più rigoroso sugli effetti di un'azione militare.

È importante notare che le difficoltà nel soddisfare il requisito di forza minima sono legate ai contesti concreti (cioè, alla possibilità per il combattente di formulare un giudizio situazionale) e non al principio di necessità stesso. In teoria, quelle difficoltà possono essere superate se vengono previste misure per mitigare i rischi generati dal problema della predicibilità. Per esempio, si potrebbe definire e far rispettare una soglia di rischio per un livello accettabile di impredicibilità degli AWS (vedi il capitolo 4 e il prossimo paragrafo). Anche trattare decisori, combattenti umani e AWS come agenti di un sistema ibrido, anziché come agenti che operano rispettando una catena di comando, può essere utile a tal fine (Taddeo et al., 2022). In questo modo si potrebbero definire protocolli di addestramento per gli agenti umani per produrre livelli appropriati di fiducia nella tecnologia, mantenere la *situational awareness* e affrontare i possibili esiti non previsti, il tutto per migliorare il controllo umano degli AWS. Si potrebbero inoltre definire standard tecnici che specifichino

opportune interfacce umani-macchine e identifichino i contesti d'uso di AWS non letali, per esempio ambienti sottomarini, dove i rischi legati al problema della predicibilità avrebbero un impatto limitato.

Affronterò nuovamente la questione nel paragrafo successivo; prima, tuttavia, desidero riprendere le argomentazioni contro la rilevanza del principio di necessità nel dibattito sugli AWS. Spero che l'analisi presentata fino a questo punto abbia convinto chi legge dell'importanza di tale principio, ma vorrei aggiungere due ulteriori considerazioni a sostegno di questa posizione. La prima riguarda l'ambito del dibattito, che finora si è focalizzato esclusivamente sui LAWS, senza considerare gli usi non letali degli AWS. Questo costituisce un punto problematico, poiché gli usi non letali degli AWS sollevano questioni etiche che non vengono affrontate nelle discussioni riguardanti i LAWS. Le discussioni sui LAWS si concentrano infatti sulle uccisioni, in particolare sulla questione se sia moralmente e legalmente consentito delegare alle macchine la decisione di uccidere, trascurando problemi cruciali legati all'intenzionalità d'uso, al controllo, alla coercizione e all'autonomia, che emergono con maggiore evidenza quando si esamina l'impiego non letale degli AWS. È probabile che gli AWS, in tale contesto, possano risultare più ammissibili dei LAWS, e per questo è fondamentale includerli nel dibattito, affrontando in modo appropriato i rischi etici connessi. Il secondo punto è legato all'obiettivo di questo capitolo, cioè identificare un terreno comune fra le opposte posizioni del dibattito sugli AWS per sfuggire alla polarizzazione continua della discussione. L'identificazione di compromessi fra necessità militare e forza minima per gli usi non letali degli AWS può orientare anche il dibattito sui LAWS. Per dirla in altro modo: se non è possibile utilizzare gli AWS in modo non letale rispettando il requisito della forza minima, allora *nessun* uso degli AWS, letale o no, può essere considerato ammissibile sul piano morale. Se invece si possono identificare compromessi accettabili che giustifichino usi non letali, questo potrebbe gettare luce anche sulle condizioni di ammissibilità dell'uso dei LAWS.

8.4 DISTINZIONE, DOPPIO EFFETTO E *DUE CARE*

Il principio di distinzione impone alle parti in un conflitto armato di distinguere fra obiettivi militari e civili, e di rivolgere le proprie azioni solo verso i primi. I combattenti sono esposti agli attacchi, perché hanno rinunciato al diritto di non essere attaccati. L'immunità dei non combattenti, invece, è assoluta, nella Teoria della Guerra Giusta (Walzer, 1977, p. 151). Va detto che il principio di distinzione ha numerose interpretazioni (Bica, 1998; Kasher, 2007). Nell'interpretazione minimale, i combattenti non devono prendere intenzionalmente a bersaglio i non combattenti. Il principio di distinzione in sé, però, "non rende illecito che i civili muoiano in tempo di guerra" (Orend, 2019, p. 112), perché permette un danno non intenzionale ma prevedibile subito dai non combattenti, se quel danno è proporzionato agli obiettivi che l'attacco intende raggiungere. Questa è la dottrina del "doppio effetto", che dà voce all'intuizione morale che

> è ammissibile causare un danno come effetto collaterale (o "doppio effetto") dell'ottenimento di un buon risultato anche se non sarebbe ammissibile causare quel danno come mezzo per realizzare quel medesimo fine buono. (McIntyre, 2004)

Walzer ha formulato le quattro condizioni della dottrina del doppio effetto in guerra, e tutte devono essere soddisfatte perché il danno a non combattenti sia ammissibile. Sono:

1. l'atto in sé è buono o almeno indifferente, il che vuol dire […] che è un atto legittimo di guerra;
2. l'effetto diretto è moralmente accettabile – la distruzione di rifornimenti militari, per esempio, o l'uccisione di soldati nemici;
3. l'intenzione dell'attore è buona, cioè, egli mira soltanto all'effetto accettabile; l'effetto cattivo non costituisce uno dei suoi fini, né rappresenta un mezzo per raggiungere tali fini;
4. l'effetto buono lo è a sufficienza per compensare il fatto di aver provocato l'effetto cattivo; esso deve essere giustificabile nei termini della regola della proporzionalità. (Walzer, 1977, p. 195)

Blanchard e io (2022b) abbiamo sostenuto che queste condizioni danno un'interpretazione minimalista del principio di distinzione, che non proibisce l'uso di AWS. È possibile che un combattente usi gli AWS con buone intenzioni, mirando a un effetto accettabile. Si può sostenere che l'impredicibilità degli AWS comporti l'imprevedibilità di un danno (non intenzionale) per i non combattenti, e se quell'effetto è militarmente buono "a sufficienza per compensare il fatto di aver provocato l'effetto cattivo" (Walzer, 1977, p. 195) in base alla regola della proporzionalità, allora un danno del genere è ammissibile.

Tuttavia, l'interpretazione minimalista non prende in considerazione l'obbligo di *due care* ("dovuta attenzione") implicato dal principio di distinzione (*ibidem*, pp. 193-202). Questo obbligo è fondamentale nel caso degli AWS (Blanchard, Taddeo, 2022b), perché definisce i rischi e gli obblighi dei combattenti nell'uso della forza (letale). Per la *due care*, i combattenti sono obbligati ad accettare rischi maggiori per sé stessi, per assicurarsi di colpire solo il bersaglio giusto e diminuire i rischi per i non combattenti (Orend, 2001, pp. 12-13).[5] La natura della *due care* non è stata esplicitata a pieno nella Teoria della Guerra Giusta, ma per molti è un aspetto centrale della condotta per lo *jus in bello*. Come scrive McMahan:

> La concezione dominante nella tradizione della guerra giusta […] è che, quando i combattenti devono scegliere fra imporre un certo rischio ai civili, come effetto collaterale delle loro azioni, e accettare un rischio ancora maggiore per sé stessi, devono, almeno fino a un certo punto, accettare il rischio maggiore. (2010, p. 344)

In una guerra, i combattenti sono responsabili del danno per la loro capacità di colpire. Ne segue che, se un combattente è responsabile del danno causato da un attacco intenzionale, la sua responsabilità deve comprendere anche quella per i danni derivanti dagli effetti collaterali dell'azione militare. Di conseguenza, rientra nel ruolo di un combattente accettare un rischio maggiore (entro i limiti definiti dal principio di necessità e dal requisito della forza minima) ed evitare di imporlo a chi non è responsabile (Margalit, Walzer, 2009). La *due care* afferma la buona intenzione evidenziata dalla dottrina del doppio effetto, attraverso la dimostrazione di moderazione in guerra (Orend, 2001, p. 13) e richiede che il danno prevedibile causato da un'azione militare sia ridotto nella misura maggiore possibile. Pertanto, rende più restrittivi la dottrina del

doppio effetto e il principio di distinzione. Prendendo in considerazione la *due care*, il principio di distinzione è soddisfatto se

> l'intenzione dell'attore è buona, cioè, egli mira esclusivamente all'effetto accettabile; l'effetto cattivo non è uno dei suoi fini, né il mezzo per raggiungere i suoi fini e, consapevole del danno che può arrecare, cerca di minimizzarlo, accettando di pagarne personalmente i costi. (Walzer, 1977, p. 198)

Quindi, il rispetto del principio di distinzione dipende dalla possibilità di distribuire in modo appropriato i rischi di danni fra combattenti e non combattenti, e anche dalla soglia oltre la quale i combattenti non sono più obbligati ad assumersi un rischio aggiuntivo (*ibidem*, p. 200).[6]

Per soddisfare entrambe le condizioni è necessaria una valutazione situazionale che dia considerazione "alla natura dell'obiettivo, all'urgenza del momento, alla tecnologia disponibile, e così via" (*ibidem*, p. 199). Per quanto riguarda gli AWS, il problema è che una valutazione di questo genere non è possibile, dati il problema della predicibilità e le condizioni estremamente dinamiche della guerra contemporanea. In effetti, il problema della predicibilità motiva la raccomandazione dell'ICRC:

> I sistemi d'arma autonomi *impredicibili* devono essere espressamente esclusi, specificamente per i loro effetti indiscriminati. Questo si può ottenere al meglio con una proibizione dei sistemi d'arma autonomi progettati o usati in un modo che non permetta di comprenderne, prevederne e spiegarne a sufficienza gli effetti. (International Committee of the Red Cross, 2021, p. 2, corsivo mio)

Tuttavia, questa posizione presuppone una dicotomia un po' ingenua fra AWS predicibili e impredicibili, dato il riferimento a "sistemi d'arma autonomi *impredicibili*", e rischia di portare a un approccio a due livelli alla regolamentazione, che non riesce a cogliere la natura più articolata del problema della predicibilità e delle sue ampie ramificazioni. Se la si limita agli aspetti tecnici degli AWS, questa distinzione è corretta in linea di principio. Esistono modelli di AWS come i modelli offline, che, considerati isolatamente, non presentano necessariamente un problema di predicibilità. Come abbiamo visto nel capitolo 1, però, i sistemi autonomi che apprendono, che siano o no prevedibili per il modo in cui sono progettati, possono comunque presentare comportamenti imprevisti quando sono in uso. La limitata predicibilità degli AWS non dipende dal loro stato di sviluppo corrente o dalle condizioni d'uso; tutti i sistemi IA sono imprevedibili, in una certa misura, a causa della loro autonomia,

delle capacità di apprendimento e della limitata robustezza. Per questo l'impredicibilità di un sistema IA aumenta con il suo livello di sofisticazione e con le complessità delle condizioni conflittuali. Come notava l'UNIDIR:

> Quanto più complesso è l'ambiente operativo in cui è utilizzato un sistema, tanto più è probabile che il sistema incontri input per i quali non è stato specificamente addestrato o testato o che presenti nuovi comportamenti che non sono stati osservati o validati in precedenza. (Holland Michel, 2020b, p. 7)

Se consideriamo questi aspetti, è evidente che non esiste una distinzione netta fra AWS predicibili e impredicibili; bisogna concentrarsi invece su *livelli di predicibilità*, sui rischi relativi e sulla specificazione di quali (se esistono) passi del processo di esercizio della forza possano ammettere l'uso di AWS, dato il loro livello di predicibilità, i rischi che generano e l'obbligo della *due care*. Analizzo questi punti nel prossimo paragrafo.

8.4.1 AWS, distinzione e *due care*

Al livello tattico dell'esecuzione di una missione, la decisione di ingaggiare il bersaglio ed esercitare la forza è preceduta da una serie di passi, fra cui individuare, fissare e tracciare il bersaglio. La *due care* impone ai combattenti di determinare l'ammissibilità del bersaglio sulla base della loro valutazione situazionale. Se si delega questo compito a un AWS, si corre un forte rischio di violare l'obbligo della *due care*, dati i limiti di questi sistemi già descritti nel paragrafo precedente.

Per quanto riguarda gli AWS e la loro capacità di identificare bersagli legittimi, la letteratura fa riferimento a casi in cui un AWS *vede* oggetti o individui e li identifica come bersagli. Nella guerra contemporanea, però, è improbabile che la scelta di un bersaglio da attaccare possa essere determinata sulla base di caratteristiche puramente visuali, per esempio il fatto di indossare un'uniforme. I marcatori visuali funzionavano all'epoca in cui la guerra veniva combattuta fra soldati in campo a una certa distanza dai centri civili (Nurick, 1945). Oggi, le attività belliche coinvolgono anche combattenti senza uniforme (per esempio nelle insurrezioni urbane) e richiedono un uso molto più ampio della *situational awareness* per determinare se un agente può essere oggetto di un attacco, perché la valutazione dipende dal comportamento del bersaglio e spesso le

differenze comportamentali sono minime. Per esempio, civili che usano un'arma per proteggere la propria famiglia, e non per ottenere un qualche vantaggio nel conflitto, non possono essere esposti a un attacco.

Classificazioni come combattenti/non combattenti e soldati/civili introducono situazioni quasi ideali, che nella realtà non si presentano, perché lo status di un agente varia con il suo comportamento. Per questo la raccomandazione di usare gli AWS solo per colpire "oggetti che sono obiettivi militari per loro natura" (International Committee of the Red Cross, 2021, p. 2) è problematica, perché si può determinare se un attore o un oggetto è un obiettivo militare solamente all'interno di un contesto specifico. Per esempio, un combattente che si è arreso o è stato neutralizzato a causa di una ferita diventa *hors de combat* e non può essere attaccato. Si potrebbe rispondere che esistono artefatti, come i carri armati, che sono oggettivamente strumenti bellici, e costituiscono essenzialmente bersagli militari, indipendentemente dal contesto d'uso, ma non è sempre così (Blanchard, Taddeo, 2022b). Per esempio, si pensi alla

> cosiddetta "Autostrada della Morte" nella prima Guerra del Golfo (Mueller, 1995). Nella notte del 26 febbraio 1991, un aereo della coalizione attaccò e distrusse una colonna di centinaia di veicoli militari iracheni con equipaggio a bordo. All'epoca, la controversia sull'incidente verteva sul fatto che le forze irachene si stessero arrendendo o stessero ritirandosi per riorganizzarsi. Nel primo caso, gli attacchi lanciati dalle forze della coalizione sarebbero stati potenzialmente in violazione della Convenzione di Ginevra (Hersh, 2000). Questa disputa testimonia che la nozione di oggetti che sono obiettivi militari per loro natura è errata. (Blanchard, Taddeo, 2022b, p. 19)

Quindi, determinare se qualcosa può essere o no attaccabile, entro uno specifico contesto, richiede una valutazione basata su una *situational awareness* ben articolata di quel dato contesto. Non deve essere una valutazione del bersaglio in quanto essere umano, e non richiede un apprezzamento della dignità e un riconoscimento del rispetto dovuto a un altro essere umano, ma deve essere una valutazione del bersaglio da colpire che sia accurata e certa, almeno quanto lo sarebbe se effettuata da un agente umano che operasse nelle medesime circostanze. In linea di principio, non è impossibile che un AWS conduca una valutazione di questo genere, ma è improbabile che raggiunga quel livello di precisione e di certezza, almeno nel prossimo futuro.

I limiti degli AWS nell'identificazione di bersagli legittimi, evidenziati in questo paragrafo, motivano la raccomandazione dell'ICRC di mettere al

bando questi sistemi d'arma, perché rischiano di danneggiare tutti quelli che sono coinvolti in un conflitto armato. Credo però che la distinzione e la *due care* si debbano considerare insieme alla necessità e alla proporzionalità, e anche insieme a una considerazione della necessità militare e dei vantaggi operativi che gli AWS potrebbero offrire. Nel loro insieme, questi fattori richiedono una posizione più articolata.

Tutti i sistemi d'arma, autonomi o no, creano rischi per tutti coloro che sono coinvolti nei conflitti armati. L'interrogativo è se quei rischi siano accettabili sul piano morale (e giuridico). Per affrontare questo interrogativo, è fondamentale determinare soglie di rischio al di sopra delle quali i rischi posti dagli AWS non sono accettabili, nonché definire misure di mitigazione del rischio, come test, addestramento, validazione del software e misure di cybersicurezza. Il consenso su questi punti si sta ampliando, fra gli esperti e i decisori politici. Per esempio, lo UN GGE CCW (2019) afferma: "Valutazioni del rischio e misure di mitigazione devono fare parte della progettazione, dello sviluppo, del test e dell'uso delle tecnologie emergenti in qualsiasi sistema d'arma". Sono convinta che concentrarsi sul rischio e le soglie di rischio sia un approccio percorribile per evitare la polarizzazione del dibattito sull'ammissibilità degli AWS e per definire un approccio più fruttuoso per colmare il vuoto di regolamentazione e identificare casi limite per l'uso di queste armi.

Bisogna stare attenti, però, perché la focalizzazione sulla gestione del rischio può diventare una scorciatoia per aggirare la riflessione sulla Teoria della Guerra Giusta e sull'etica della guerra, che deve stare alla base di qualsiasi approccio di regolamentazione dell'uso degli AWS. È cruciale che il dibattito sull'ammissibilità degli AWS non si riduca a una semplice questione di gestione del rischio, trascurando i fattori normativi che devono orientare la distribuzione dei rischi e la definizione di soglie di rischio accettabili, di cui abbiamo parlato in questo capitolo.

Entrambi i principi, di necessità e di distinzione, si basano sull'allocazione del rischio e sulla risoluzione del compromesso fra protezione della forza e minimizzazione del danno in guerra sia per i combattenti sia per i non combattenti (McMahan, 2010, p. 343). Le risposte non sono neutre rispetto ai valori; sono determinate da scelte morali e politiche (Perry, 1995; Garland, 2003). Pertanto, la definizione delle soglie e delle misure di mitigazione dei rischi richiede un dibattito pubblico, un forum multistakeholder e un terreno che renda possibile la

discussione. La Teoria della Guerra Giusta offre questo terreno; quello che serve è un modello per la distribuzione dei rischi coerente con i principi di questa teoria. Senza un terreno normativo solido come questo, gli approcci di valutazione del rischio per gli AWS non risolveranno il dibattito sulla loro ammissibilità.

8.5 CONCLUSIONE

Obiettivo di questo capitolo è stato offrire un'interpretazione della Teoria della Guerra Giusta che possa costituire un terreno comune per le diverse posizioni nel dibattito sull'ammissibilità degli AWS, nella speranza di superare l'attuale polarizzazione e affrontare il vuoto di regolamentazione sull'uso di queste armi. L'analisi dei principi di necessità e di distinzione mostra che, se ci si concentra sulle soglie di rischio, è possibile identificare casi limite nell'uso degli AWS senza basarsi eccessivamente sull'elemento consequenzialista o su quello deontologico della Teoria della Guerra Giusta e tenendo conto, al contempo, della necessità militare.

Quest'analisi porta a tre conclusioni fondamentali. La prima è che il rispetto dei principi di necessità e distinzione richiede un giudizio e una *situational awareness* adeguati a valutare scenari di combattimento e rischi correlati, e a prendere decisioni strategiche e tattiche giustificabili. Non vale nemmeno la pena di porsi la domanda se agli AWS possa essere demandata una simile valutazione nel rispetto di quei principi. Da un lato, l'idea che questi sistemi possano sviluppare un livello appropriato di *situational awareness* per affrontare le complessità degli scenari di guerra contemporanei non è realistica; dall'altro, i limiti di predicibilità e robustezza indicano che gli AWS possono presentare comportamenti diversi da quelli che ci si aspettava, e di conseguenza è grave il rischio che possano violare i principi che invece dovrebbero rispettare.

La seconda conclusione è che gli AWS possono dare vantaggi tattici e strategici importanti, che possono giustificarne l'uso, ma, in base all'analisi proposta qui, quell'uso va pensato come complementare, anziché alternativo, ai combattenti umani. Gli AWS vanno impiegati all'interno di un team ibrido, in cui collaborino con gli umani; in tal caso, gli interrogativi sull'ammissibilità degli AWS si possono risolvere affrontando problemi relativi al contesto in cui sia accettabile costituire team umani-macchine in guerra, quali compiti si possano delegare agli AWS, quale livello di autonomia debbano avere, quale tipo di addestramento sia necessario per i combattenti che lavorano con gli AWS, come progettare interfacce per migliorare la trasparenza e il controllo

sugli AWS, e come identificare, valutare e mitigare i rischi relativi, e definire soglie di rischio e standard per valutare la predicibilità degli AWS (Tsamados, Taddeo, 2023).

La terza conclusione è che l'uso degli AWS in modalità completamente autonoma rimane inaccettabile sul piano morale, perché quell'uso genererebbe un rischio troppo grave di violare i principi della Teoria della Guerra Giusta. In ultima istanza, qualsiasi uso degli AWS deve essere regolato sicuramente tenendo conto della necessità militare ma, cosa molto più importante, tenendo conto di ciò che è accettabile nel rispetto dei valori e dei diritti che ne sono alla base.

---

1. Un chiarimento è necessario prima di iniziare l'analisi: nel resto del capitolo farò riferimento agli AWS come sono stati definiti nel capitolo 6; quindi mi concentrerò sugli agenti artificiali autonomi dotati di capacità di apprendimento che consentono di adattare il loro comportamento alle condizioni di utilizzo e che servono all'applicazione della forza, letale o no. Userò AWS per riferirmi ai sistemi d'arma autonomi sia letali sia non letali. Specificherò quando non sarà questo il caso e l'analisi verterà su un sottoinsieme specifico di AWS, per esempio LAWS o AWS non letali.

2. Gran parte dell'opposizione agli AWS sulla base del principio di ultima istanza è analoga alle preoccupazioni che sono state sollevate per l'uso bellico dei droni (vedi Strawser, 2013). Per chiarezza, qui mi concentro sugli AWS.

3. La confusione generata dall'aver trascurato tali fattori ha una lunga storia ed è emersa in modo piuttosto evidente nel primo impiego dei droni in contesti di combattimento, e nell'argomentazione, sostenuta dai loro promotori, secondo cui, se considerati lungo una "linea di tendenza storica", i droni possono essere interpretati come strumenti in grado di generare guerre più proporzionate (K. Anderson, s.d., pp. 383-384). Tuttavia, spesso questa "linea di tendenza storica" prendeva come riferimento la Seconda guerra mondiale e, dato che le forze alleate colpivano intenzionalmente la popolazione civile, sembrava poco lodevole usare come parametro di confronto per i droni quello che può essere considerato un "punto più basso storico della guerra" (Braun, Brunstetter, 2013, p. 309; vedi anche Shue, 2008).

4. Centro studi per la pace: https://www.studiperlapace.it/view_news_html?news_id=20041101105835.

5. La *due care* è stata codificata nella Law of Armed Conflict in quanto precauzione nell'attacco (Ministry of Defence, 2011, pp. 81-88).

6. Per obiettivo militare *giustificato* si intende che l'obiettivo sia conforme agli altri principi dello *jus in bello*, come il principio di necessità e proporzionalità.

EPILOGO

Il cielo fa con noi come noi con le torce,
che non s'accendono solo per sé stesse:
se dalle nostre virtù non irradia luce,
tanto varrebbe non averle.

WILLIAM SHAKESPEARE,
Misura per misura, atto I, scena I

Ho finito di scrivere la versione inglese del libro a quasi due anni dall'inizio della guerra in Ucraina e a pochi mesi dall'inizio del conflitto a Gaza. Sullo sfondo di questi eventi laceranti, non ho potuto fare a meno di domandarmi se un libro come questo avesse, in definitiva, qualche utilità. Un'etica della guerra può essere teoricamente giustificata e moralmente necessaria, ma rischia di essere irrilevante se gli Stati e i loro leader politici ne ignorano sistematicamente le prescrizioni. Sarebbe ingenuo non domandarsi a cosa serva un'etica della difesa, quando nella prima guerra combattuta in Europa dopo la Seconda guerra mondiale entrambi gli schieramenti usano AWS prima che il dibattito etico sulla loro legittimità sia concluso, e numerose voci, dalla società civile alle istituzioni sovranazionali, ne invocano il divieto. L'uso di sistemi IA per identificare bersagli umani a Gaza, così come riportato – privo di una supervisione umana adeguata e con alte soglie di errore –, e le conseguenti violazioni dei principi di proporzionalità, distinzione e necessità, sembrano a loro volta testimoniare l'irrilevanza di un'etica dell'IA nella difesa.

Dopo un iniziale momento di scoramento, la cruda realtà di questi conflitti mi ha spinto, se non altro, a tentare di fare la differenza. Come ha scritto Walzer (2006), "è dalle limitazioni della guerra che germogliano i semi della pace". Ne segue la necessità di contenere e governare l'uso delle tecnologie digitali – e in particolare dell'IA – quando impiegate nella difesa.

Nei mesi precedenti la pubblicazione della versione in italiano (agosto 2025), il "tentativo di fare la differenza" si è trasformato in urgenza, dettata dal ritorno alla cosiddetta *gunboat diplomacy*, la diplomazia della cannoniera. Questa forma anacronistica di diplomazia sfida il principio alla base del diritto internazionale – l'interdizione agli Stati di annettere *manu militari* territori appartenenti a paesi terzi. Eliminando il diritto di conquista, il diritto internazionale ha spinto gli Stati a perseguire la crescita economica attraverso mezzi pacifici, facendo leva soprattutto su incentivi e sanzioni economiche. Come riportano Hathaway e Shapiro (2018), la trasformazione è stata tanto radicale quanto misurabile: nei sessantacinque anni successivi ai trattati di pace della Seconda guerra mondiale, la quantità di territorio conquistato da Stati stranieri ogni anno è precipitata a meno del 6% di quello che era stato per poco più di un secolo prima che il mondo mettesse fuorilegge la guerra per la prima volta.

Il diritto internazionale nasce con una transizione dalla coercizione veicolata con la forza alla coercizione diplomatica, mediata da istituzioni sovranazionali e codici normativi condivisi. Componente cruciale di questa transizione è stata l'elaborazione di una disciplina minuziosa del diritto internazionale umanitario (IHL), che governa con precisione ogni aspetto dei processi decisionali e operativi delle attività belliche. L'obiettivo è definire i criteri per valutare la legittimità dei conflitti e l'ammissibilità delle prassi che ne seguono. In questo senso, il diritto internazionale e l'IHL in particolare non sono solo un insieme di regole e burocrazia, ma denotano l'impegno a evitare la guerra e a preservare un residuo di moralità e liceità quando questa diventa ineluttabile.

Il ritorno della *gunboat diplomacy* è un tentativo di invertire la transizione verso un ordine internazionale stabile e basato su regole condivise e di ignorare l'impegno a evitare la guerra o comunque a rispettare valori fondamentali quando la guerra si rende necessaria. Ne sono un esempio l'invasione dell'Ucraina per annetterne una parte alla Russia, il conflitto a Gaza e le dichiarazioni sull'annessione della Groenlandia agli Stati Uniti. Le conseguenti posture muscolari e unilaterali degli Stati che ricorrono a questa forma di diplomazia sfidano apertamente gli istituti del diritto internazionale. In questo scenario, l'uso dell'IA nella difesa può rivelarsi uno strumento tanto efficace quanto discreto per erodere il diritto internazionale dall'interno e ridurlo a un mero fantoccio. L'impiego di sistemi IA nelle operazioni di difesa – in

assenza di una regolamentazione adeguata – introduce una zona grigia in cui gli Stati possono eluderne sistematicamente i principi senza incorrere in sanzioni formali.

Considerato l'ingente impegno profuso nell'adozione dell'IA nella difesa, questa zona grigia risulta vantaggiosa tanto per gli Stati liberali quanto per quelli autoritari. Entrambi possono beneficiare dell'assenza di vincoli normativi chiari per sviluppare e usare sistemi IA nella difesa perseguendo i propri obiettivi strategici senza dover rendere conto alle istituzioni internazionali. Tale convergenza di interessi favorisce, paradossalmente, il disinteresse degli Stati liberali per i processi di governance dell'IA nella difesa, compromettendo la loro credibilità nel promuovere un ordine internazionale basato sullo Stato di diritto. Questa dinamica rischia di trasformare l'innovazione tecnologica da strumento di progresso in catalizzatore di regressione normativa, minando le fondamenta stesse del sistema internazionale costruito nel secondo dopoguerra. La convergenza tra il revival della *gunboat diplomacy* e il vuoto regolamentativo che circonda la trasformazione della difesa innescata dall'IA configura uno scenario preoccupante: è un'opportunità strategica per quegli attori che percepiscono il diritto internazionale come un ostacolo e non come una garanzia di stabilità sistemica.

È una convergenza preoccupante anche per il futuro delle democrazie liberali, perché esiste una relazione di reciproca influenza tra il modo in cui i conflitti vengono condotti e le società che li combattono. Come osservava Clausewitz, più che un'arte o una scienza, la guerra è un'attività sociale; come gran parte delle altre attività sociali, i conflitti rispecchiano i valori delle società e sfruttano i loro sviluppi tecnologici e scientifici. A loro volta, i principi che utilizziamo per regolare la condotta in guerra hanno un ruolo fondamentale nel plasmare le nostre società. Basti pensare alla progettazione, all'uso e alla regolamentazione delle armi di distruzione di massa: nel corso della Seconda guerra mondiale, queste armi sono state rese possibili da svolte nel campo della fisica nucleare. La violenza catastrofica che è stata scatenata sulle città di Hiroshima e Nagasaki però ha portato a un consenso globale, mai visto fino ad allora, che ha determinato nel dopoguerra l'avversione del mondo per l'uso di quelle armi. La Guerra fredda e i trattati nucleari che vi hanno posto fine hanno definito le modalità in cui le tecnologie nucleari e le armi di distruzione di massa possono essere utilizzate, tracciando un confine tra

conflitti e atrocità. In questo modo, i trattati e le regolamentazioni per l'uso delle armi di distruzione di massa hanno contribuito a orientare le società contemporanee verso il rifiuto della retorica bellicosa degli inizi del XX secolo, per tendere invece alla pace e alla stabilità.

Le cose non sono diverse oggi con le società digitali. Da un lato, l'uso dell'IA nella difesa ha un grande potenziale di migliorare il funzionamento delle organizzazioni della difesa, di rafforzarne le capacità e di rendere le operazioni militari più sicure, efficaci ed efficienti. Dall'altro, sappiamo che, se rimane non regolato, l'uso dell'IA nella difesa può avere implicazioni negative gravi in termini di stabilità (come abbiamo visto nel capitolo 4), di diritti degli individui e dei gruppi, di violazioni dei principi della Teoria della Guerra Giusta (come ho mostrato nei capitoli 3-8). Si pensi, per esempio, al caso dell'intelligence open source nel conflitto russo-ucraino. L'uso diffuso degli smartphone fra i cittadini ha consentito al personale militare di sfruttare intelligence raccolta dalla popolazione civile e condivisa sui social media per ottenere stime approssimate della posizione dei combattenti nemici. Questo ha sollevato preoccupazioni crescenti per l'estensione della sorveglianza militare della società civile e per il rischio di coinvolgere la popolazione civile nelle operazioni militari. Qui la domanda è se siamo disposti ad accettare che il *trade-off* tra efficienza della difesa e protezione dei civili sia a favore della prima.

Un quadro etico sull'uso dell'IA nella difesa deve essere fermo riguardo ai rischi etici e ai casi limite, ma allo stesso tempo deve essere capace di identificare il potenziale positivo dell'IA per la difesa e offrire linee guida per sfruttarlo in conformità con i valori che fondano le nostre società. Questo equilibrio è delicato e particolarmente complesso. Esige un sostanziale investimento di tempo e risorse, e richiede uno sforzo congiunto da parte di studiosi, *tech providers*, decisori politici e operatori del settore della difesa. Gli studiosi, in particolare gli eticisti, svolgono un ruolo cruciale nell'individuare valori, principi, teorie del valore e persino linee guida per implementare i loro quadri di riferimento. Considerato lo scenario geopolitico attuale e il ritmo di adozione dell'IA nella difesa, questo lavoro è estremamente urgente. Tuttavia, la maggior parte della responsabilità e del carico ricade sugli attori statali e sull'apparato della difesa: devono affrontare le sfide etiche dell'IA e adottare quadri etici adeguati, robusti e sviluppati in modo indipendente per affrontarle in modo autentico. Ciò implica accettare che la governance etica dell'IA nella

difesa debba essere tanto completa e profonda quanto i cambiamenti che questa tecnologia comporta. Alla fine, questa è l'unica via per garantire che l'IA funzioni come una tecnologia per la stabilità e, possibilmente, per la pace, per evitare che sia un mero strumento di guerra e che finisca per facilitare il ritorno di posture anacronistiche e antitetiche ai valori che, a caro prezzo, abbiamo eletto a fondamento delle democrazie liberali dopo la fine della Seconda guerra mondiale.

(Luglio 2025)

BIBLIOGRAFIA

ABNEY, KEITH (2013), "Autonomous robots and the future of Just War Theory". In FRITZ ALLHOFF, NICHOLAS G. EVANS, ADAM HENSCHKE (a cura di), *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century*. Routledge, London, pp. 338-351.

"ACALVIO AUTONOMOUS DECEPTION" (2019). Acalvio. https://www.acalvio.com/. Ultimo accesso luglio 2024.

ADENEY, DOUGLAS, JOHN WECKERT (1997), *Computer and Information Ethics*. Praeger, Westport, CT.

AKHGAR, BABAK, SIMEON YATES (2013), *Strategic Intelligence Management: National Security Imperatives and Information and Communications Technologies*. Elsevier/Butterworth-Heinmann, Waltham, MA.

ALEXY, ROBERT (2002), *A Theory of Constitutional Rights*. Oxford University Press, New York.

ALJUNIED, SYED MOHAMMED AD'HA (2020), "The securitization of cyberspace governance in Singapore". In *Asian Security*, 16 (3), pp. 343-362. https://doi.org/10.1080/14799855.2019.1687444.

ALLENBY, BRADEN (2013), "Emerging technologies and Just War Theory". In FRITZ ALLHOFF, NICHOLAS G. EVANS, ADAM HENSCHKE (a cura di), *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century*. Routledge, London, pp. 289-300.

ALSHAMMARI, MAJED, ANDREW SIMPSON (2017), "Towards a principled approach for engineering privacy by design". In ERICH SCHWEIGHOFER, HERBERT LEITOLD, ANDREAS MITRAKAS, KAI RANNENBERG (a cura di), *Privacy Technologies and Policy*. Springer International Publishing, Cham, pp. 161-177. https://doi.org/10.1007/978-3-319-67280-9_9.

ALSTON, PHILIP (2010), "Report of the special rapporteur on extrajudicial, summary or arbitrary executions, Philip Alston: Addendum – Study on targeted killings (A/HRC/14/24/Add.6) – Russian Federation". In *ReliefWeb*, 28 maggio. https://reliefweb.int/report/russian-federation/report-special-rapporteur-extrajudicial-summary-or-arbitrary-executions.

AMOROSO, DANIELE, GUGLIELMO TAMBURRINI (2020), "Autonomous weapons systems and meaningful human control: Ethical and legal issues". In *Current Robotics Reports*, 1 (4), pp. 187-194. https://doi.org/10.1007/s43154-020-00024-3.

ANDERSON, ANDREW, JONATHAN DODGE, AMRITA SADARANGANI, ZOE JUOZAPAITIS, EVAN NEWMAN, JED IRVINE, SOUTI CHATTOPADHYAY, MATTHEW OLSON, ALAN FERN, MARGARET BURNETT (2020), "Mental models of mere mortals with explanations of reinforcement learning". In *ACM Transactions on Interactive Intelligent Systems*, 10 (2), pp. 1-37. https://doi.org/10.1145/3366485.

ANDERSON, DAVID (2016), *Report of the Bulk Powers Review*. Independent Reviewer of Terrorism Legislation, London. https://terrorismlegislationreviewer.independent.gov.uk/wp-

content/uploads/2016/08/Bulk-Powers-Review-final-report.pdf.
ANDERSON, KENNETH (2012), "Efficiency *in bello* and *ad bellum*: Making the use of force too easy?". In CLAIRE FINKELSTEIN, JENS DAVID OHLIN, ANDREW ALTMAN (a cura di), *Targeted Killings: Law and Morality in an Asymmetrical World*. Oxford University Press, New York, pp. 374-399.
ANDERSON, KENNETH, DANIEL REISNER, MATTHEW C. WAXMAN (2014), "Adapting the law of armed conflict to autonomous weapon systems". In *International Law Studies*, 90, pp. 386-411.
ANDRAS, PETER, LUKAS ESTERLE, MICHAEL GUCKERT, THE ANH HAN, PETER R. LEWIS, KRISTINA MILANOVIC, TERRY PAYNE, ET AL. (2018), "Trusting intelligent machines: Deepening trust within socio-technical systems". In *IEEE Technology and Society Magazine*, 37 (4), pp. 76-83. https://doi.org/10.1109/MTS.2018.2876107.
ARKIN, RONALD (2009), "Ethical robots in warfare". In *IEEE Technology and Society Magazine*, 28 (1), pp. 30-33.
ARKIN, RONALD (2018), "Lethal autonomous systems and the plight of the non-combatant". In RYAN KIGGINS (a cura di), *The Political Economy of Robots*. Spring, Cham, pp. 317-326.
ARQUILLA (1999), "Ethics and information warfare". In ZALMAY KHALILZAD, JOHN PATRICK WHITE (a cura di), *Strategic Appraisal: The Changing Role of Information in Warfare*. RAND, Santa Monica, CA, pp. 379-401.
ARTICLE36 (2018), "Shifting definitions – The UK and autonomous weapons systems July 2018". http://www.article36.org/wp-content/uploads/2018/07/Shifting-definitions-UK-and-autonomous-weapons-July-2018.pdf.
ASARO, PETER (2008), "How just could a robot war be". In ADAM BRIGGLE, KATINKA WAELBERS, PHILIP A.E. BREY (a cura di), *Proceeding of the 2008 Conference on Current Issues in Computing and Philosophy*. IOS Press, Amsterdam, pp. 50-64.
ASARO, PETER (2012), "On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making". In *International Review of the Red Cross*, 94 (886), pp. 687-709. https://doi.org/10.1017/S1816383112000768.
ASARO, PETER (2020), "Autonomous weapons and the ethics of artificial intelligence". In S. MATTHEW LIAO (a cura di), *Ethics of Artificial Intelligence*. Oxford University Press, New York, pp. 212-236.
ATHALYE, ANISH, LOGAN ENGSTROM, ANDREW ILYAS, K. KWOK (2018), "Synthesizing robust adversarial examples". 7 giugno. https://www.semanticscholar.org/paper/Synthesizing-Robust-Adversarial-Examples-Athalye-Engstrom/8dce99e33c6fceb3e79023f5894fdbe733c91e92.
AYLING, JACQUI, ADRIANE CHAPMAN (2022), "Putting AI ethics to work: Are the tools fit for purpose?". In *AI and Ethics*, 2 (3), pp. 405-429. https://doi.org/10.1007/s43681-021-00084-x.
BABUTA, ALEXANDER, MARION OSWALD (2020), "Data analytics and algorithms in policing in England and Wales: Towards a new policy framework". Occasional paper. Royal United Services Institute for Defence Studies, London.
BABUTA, ALEXANDER, MARION OSWALD, ARDI JANJEVA (2020), "Artificial Intelligence and UK National Security: Policy considerations". Occasional paper. Royal United Services Institute for Defence Studies, London.
BAZARGAN, SABA (2014), "Killing minimally responsible threats". In *Ethics*, 125 (1), pp. 114-136. https://doi.org/10.1086/677023.
BEARD, JACK M. (2018), "The principle of proportionality in an era of high technology". In CHRISTOPHER M. FORD, WINSTON S. WILLIAMS (a cura di), *Complex Battlespaces: The Law of Armed Conflict and the Dynamics of Modern Warfare*. Oxford University Press, New York, pp. 261-270.

BEBBER, ROBERT (2018), "There is no such thing as cyber deterrence. Please stop". In *Cypher Brief*, 1° aprile. https://www.thecipherbrief.com/column_article/no-thing-cyber-deterrence-please-stop.

"BEHAVIOSEC: CONTINUOUS AUTHENTICATION THROUGH BEHAVIORAL BIOMETRICS" (2019). BehavioSec. https://www.behaviosec.com/.

BEKELE, ESUBE, WALLACE E. LAWSON, ZACHARY HORNE, SANGEET KHEMLANI (2018), "Human-level explanatory biases for person re-identification". In *HRI 2018: Explainable Robotic Systems*, 2, pp. 1-2.

BEKELE, ESUBE, CODY NARBER, WALLACE LAWSON (2017), "Multi-attribute Residual Network (MAResNet) for soft-biometrics recognition in surveillance scenarios". In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, Washington, DC, pp. 386-393. https://doi.org/10.1109/FG.2017.55.

BENDIEK, ANNEGRET, TOBIAS METZGER (2015), "Deterrence theory in the cyber-century: Lessons from a state-of-the-art literature review". In *Lecture Notes in Informatics* (LNI). Gesellschaft für Informatik, Bonn, pp. 553-570.

BENTHAM, JEREMY (1789), *An Introduction to the Principles of Morals and Legislation*. Doubleday, Garden City, NY (*Introduzione ai princìpi della morale e della legislazione*. Tr. it. UTET, Torino 2013).

BERGADANO, F. (1991), "The problem of induction and machine learning". In *IJCAI-91: Proceedings of the 12th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, vol. 2, pp. 1073-1078.

BERNAL, PAUL (2016), "Data gathering, surveillance and human rights: Recasting the debate". In *Journal of Cyber Policy*, 1 (2), pp. 243-264. https://doi.org/10.1080/23738871.2016.1228990.

BETZ, DAVID J., TIM STEVENS (2013), "Analogical reasoning and cyber security". In *Security Dialogue*, 44 (2), pp. 147-164. https://doi.org/10.1177/0967010613478323.

BICA, CAMILLO C. (1998), "Interpreting Just War Theory's *jus in bello* criterion of discrimination". In *Public Affairs Quarterly*, 12 (2), pp. 157-168.

BIGGIO, BATTISTA, FABIO ROLI (2018), "Wild patterns: Ten years after the rise of adversarial machine learning". In *Pattern Recognition*, 84 (dicembre), pp. 317-331. https://doi.org/10.1016/j.patcog.2018.07.023.

BIOMETRICS AND SURVEILLANCE CAMERA COMMISSIONER (2017), *National Surveillance Camera Strategy for England and Wales*. Biometrics and Surveillance Camera Commissioner, Whitehall.

BLANCHARD, ALEXANDER, MARIAROSARIA TADDEO (2022a), "*Jus in Bello* necessity, the requirement of minimal force, and autonomous weapon systems". In *Journal of Military Ethics*, 21 (3-4), pp. 286-303. https://doi.org/10.1080/15027570.2022.2157952.

BLANCHARD, ALEXANDER, MARIAROSARIA TADDEO (2022b), "Predictability, distinction & due care in the use of lethal autonomous weapons systems". In *SSRN Electronic Journal*, 3 maggio. http://dx.doi.org/10.2139/ssrn.4099394.

BLANCHARD, ALEXANDER, MARIAROSARIA TADDEO (2022c), "Autonomous weapon systems and *jus ad bellum*". In *AI & Society*, marzo. https://doi.org/10.1007/s00146-022-01425-y.

BLANCHARD, ALEXANDER, MARIAROSARIA TADDEO (2023), "The ethics of artificial intelligence for intelligence analysis: A review of the key challenges with recommendations". In *Digital Society*, 2 (1), n. 12. https://doi.org/10.1007/s44206-023-00036-4.

BLANCHARD, ALEXANDER, CHRIS THOMAS, MARIAROSARIA TADDEO (2023), "Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance". In *SSRN Electronic Journal*, 21 luglio. https://doi.org/10.2139/ssrn.4517701.

BOARDMAN, MICHAEL, FIONA BUTCHER (2019), “An exploration of maintaining human control in AI enabled systems and the challenges of achieving it”. In *STO-MP-IST-178*, pp. 1-16.
BOCA, PAUL (2014), *Formal Methods: State of the Art and New Directions*. Springer, London.
BOLOGNA, SANDRO, ALESSANDRO FASANI, MAURIZIO MARTELLINI (2013), “From fortress to resilience”. In MAURIZIO MARTELLINI (a cura di), *Cyber Security: Deterrence and IT Protection for Critical Infrastructures*. Springer, Heidelberg, pp. 53-56.
BOOGAARD, JEROEN VAN DEN (2015), “Proportionality and autonomous weapons systems”. In *Journal of International Humanitarian Legal Studies*, 6 (2), pp. 247-283.
BOULANIN, VINCENT, MOA PELDÁN CARLSSON, NETTA GOUSSAC, DAVISON DAVIDSON (2020), “Limits on autonomy in weapon systems: Identifying practical elements of human control”. Stockholm International Peace Research Institute and the International Committee of the Red Cross. https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0.
BRANSCOMBE, NYLA R., SUSAN OWEN, TERI A. GARSTKA, JASON COLEMAN (1996), “Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome?”. In *Journal of Applied Social Psychology*, 26 (12), pp. 1042-1067. https://doi.org/10.1111/j.1559-1816.1996.tb01124.x.
BRAUN, MEGAN, DANIEL R. BRUNSTETTER (2013), “Rethinking the criterion for assessing CIA-targeted killings: Drones, proportionality and *jus ad vim*”. In *Journal of Military Ethics*, 12 (4), pp. 304-324.
BRENT, LAURA (2019), “NATO’s role in cyberspace”. In *NATO Review*, 12 febbraio. https://www.nato.int/docu/review/articles/2019/02/12/natos-role-in-cyberspace/index.html.
BREWSTER, THOMAS (2020), “Google promised not to use its AI in weapons, so why is it investing in startups straight out of ‘star wars’?”. In *Forbes*, 22 dicembre. https://www.forbes.com/sites/thomasbrewster/2020/12/22/google-promised-not-to-use-its-ai-in-weapons-so-why-is-alphabet-investing-in-ai-satellite-startups-with-military-contracts/.
BRITTAIN, STEPHEN (2016), “Justifying the teleological methodology of the European Court of Justice: A rebuttal”. In *Irish Jurist, new series*, 55, pp. 134-165.
BRODIE, BERNARD (1978), “The development of nuclear strategy”. In *International Security*, 2 (4), pp. 65-68.
BRUNDAGE, MILES, SHAHAR AVIN, JACK CLARK, HELEN TONER, PETER ECKERSLEY, BEN GARFINKEL, ALLAN DAFOE, ET AL. (2018), “The malicious use of artificial intelligence: Forecasting, prevention, and mitigation”. In *arXiv:1802.07228 [Cs]*, febbraio. http://arxiv.org/abs/1802.07228.
BRUNSTETTER, DANIEL, MEGAN BRAUN (2013), “From *jus ad bellum* to *jus ad vim*: Recalibrating our understanding of the moral use of force”. In *Ethics & International Affairs*, 27 (1), pp. 87-106. https://doi.org/10.1017/S0892679412000792.
BUCHANAN, BEN, ANDREW IMBRIE (2022), *The New Fire: War, Peace, and Democracy in the Age of AI*. MIT Press, Cambridge, MA.
BUNN, M.E. (2007), “Can deterrence be tailored?”. Strategic Forum, n. 225, gennaio. Institute for National Strategic Studies, National Defense University, Washington, DC.
BURGESS, MATT (2017), “What is the Petya ransomware spreading across Europe? WIRED explains”. In *Wired*, sez. *Security*. https://www.wired.com/story/petya-malware-ransomware-attack-outbreak-june-2017/.
CAMBRIDGE CONSULTANTS (2019), “Use of AI in online content moderation”. Report per Ofcom. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

CAMERON, DAVID, GREAT BRITAIN CABINET OFFICE (2010), *A Strong Britain in an Age of Uncertainty: The National Security Strategy*. TSO, London.

CAMPEDELLI, GIAN MARIA, MIHOVIL BARTULOVIC, KATHLEEN M. CARLEY (2021), "Learning future terrorist targets through temporal meta-graphs". In *Scientific Reports*, 11 (1), pp. 1-15.

CASTELFRANCHI, CRISTIANO, RINO FALCONE (2003), "From automaticity to autonomy: The frontier of artificial agents". In HENRY HEXMOOR, CRISTIANO CASTELFRANCHI, RINO FALCONE (a cura di), *Agent Autonomy*. Springer US, Boston, MA, pp. 103-136. https://doi.org/10.1007/978-1-4419-9198-0_6.

CATH, CORINNE, SANDRA WACHTER, BRENT MITTELSTADT, MARIAROSARIA TADDEO, LUCIANO FLORIDI (2018), "Artificial intelligence and the 'good society': The US, EU, and UK approach". In *Science and Engineering Ethics*, 24 (2), pp. 505-528.

CCW GGE (2019), "Guiding principles affirmed by the group of governmental experts on emerging technologies in the area of lethal autonomous weapons system (Annex III)". CCW/MSP/2019/9. United Nations Office of Disarmament Affairs, Genève. https://www.ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf.

CHAMPAGNE, MARC, RYAN TONKENS (2015), "Bridging the responsibility gap in automated warfare". In *Philosophy & Technology*, 28 (1), pp. 125-137.

CHEN, JIM Q. (2016), "Intelligent targeting with contextual binding". In *2016 Future Technologies Conference (FTC)*, pp. 1040-1046. IEEE, San Francisco, CA. https://doi.org/10.1109/FTC.2016.7821732.

CHENGETA, THOMPSON (2016), "Measuring autonomous weapon systems against international humanitarian law rules". In *Journal of Law & Cyber Warfare*, 5 (1), pp. 66-146.

CHOPRA, AMIT K., MUNINDAR P. SINGH (2018), "Sociotechnical systems and ethics in the large". In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 48-53. Association for Computing Machinery, New York. https://doi.org/10.1145/3278721.3278740.

CHOUDHURY, L., A. AOUN, D. BADAWY, L.A. DE ALBURQUERQUE, J. MARJANE, A. WILKINSON (2021), "Letter [from] the panel of experts on Libya established pursuant to Resolution 1973 (2011) addressed to the President of the Security Council". *s/2021/229. United Nations Security Council.*

CICERO, MARCUS TULLIUS (2008), *On Obligations*. Tr. ing. di Patrick G. Walsh. Oxford University Press, Oxford (*I doveri*. Tr. it. Rizzoli, Milano 1992).

CIHON, PETER, JONAS SCHUETT, SETH D. BAUM (2021), "Corporate governance of artificial intelligence in the public interest". In *Information*, 12 (7), n. 275. https://doi.org/10.3390/info12070275.

CINA (2018), "Convention on certain conventional weapons: Position paper submitted by China". https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_(2022)/CCW-GGE.1-2022-WP.6.pdf.

CLARK, DAVID, SUSAN LANDAU (2011), "Untangling attribution". In *Harvard National Security Journal*, 2011 (2), pp. 25-40.

CLARK, IAN (2015), *Waging War: A New Philosophical Introduction*. Oxford University Press, Oxford.

CLAUSEWITZ, CARL VON (1832), *On War*. Tr. ing. di James John Graham. Wilder Publications, Radford, VA, 2008 (*Della Guerra*. Tr. it. Mondadori, Milano 1978).

COLDICUTT, RACHEL, CATHERINE MILLER (2019), "People, power, and technology: The tech workers' view". Doteveryone, London. https://doteveryone.org.uk/wp-content/uploads/2019/04/PeoplePowerTech_Doteveryone_May2019.pdf.

COLEMAN, STEPHEN (2015), "Possible ethical problems with military use of non-lethal weapons international regulation of emerging military technologies". In *Case Western Reserve Journal of International Law*, 47 (1), pp. 185-200.
COLLOPY, PAUL, VALERIE SITTERLE, JENNIFER PETRILLO (2020), "Validation testing of autonomous learning systems". In *Insight*, 23 (1), pp. 48-51. https://doi.org/10.1002/inst.12285.
CONN, ARIEL (2016), "The problem of defining autonomous weapons". Future of Life Institute. 30 novembre. https://futureoflife.org/2016/11/30/problem-defining-autonomous-weapons/.
CONVENTION ON CERTAIN CONVENTIONAL WEAPONS (2018), "Report of the 2018 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems". CCW/GGE.1/2018/3. United Nations Office for Disarmament Affairs, Genève. https://documents.un.org/doc/undoc/gen/g18/323/29/pdf/g1832329.pdf. Ultimo accesso luglio 2024.
CORLETT, J. ANGELO (2001), "Collective moral responsibility". In *Journal of Social Philosophy*, 32 (4), pp. 573-584. https://doi.org/10.1111/0047-2786.00115.
CORNILLE, CHRIS (2021), "AI experts needed to lead 'project maven' move within DOD". In *Bloomberg Government* (blog), 1º giugno. https://about.bgov.com/insights/news/ai-experts-needed-to-lead-project-maven-move-within-dod/.
COVERDALE, JOHN F. (2004), "An introduction to the Just War tradition". In *Pace International Law Review*, 16 (2), pp. 221-278.
CROSSTON, MATTHEW (2011), "World gone cyber MAD: How 'mutually assured debilitation' is the best hope for cyber deterrence". In *Strategic Studies Quarterly*, 50 (1), pp. 100-116.
CUMMINGS, MARY, SONGPO LI (2019), "HAL2019-02: Machine learning tools for informing transportation technology and policy". Humans and Autonomy Laboratory, Duke University. http://hal.pratt.duke.edu/sites/hal.pratt.duke.edu/files/u39/HAL2019_2%5B1920%5D-min.pdf.
D'AQUIN, MATHIEU, PINELOPI TROULLINOU, NOEL E. O'CONNOR, AINDRIAS CULLEN, GRÁINNE FALLER, LOUISE HOLDEN (2018), "Towards an 'ethics by design' methodology for AI research projects". In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 54-59. Association for Computing Machinery, New Orleans, LA. https://doi.org/10.1145/3278721.3278765.
"DARKLIGHT OFFERS FIRST OF ITS KIND ARTIFICIAL INTELLIGENCE TO ENHANCE CYBERSECURITY DEFENSES" (2017). In *Business Wire*, 26 luglio. https://www.businesswire.com/news/home/20170726005117/en/DarkLight-Offers-Kind-Artificial-Intelligence-Enhance-Cybersecurity.
DAVIES, RACHEL, JONATHAN IVES, MICHAEL DUNN (2015), "A systematic review of empirical bioethics methodologies". In *BMC Medical Ethics*, 16 (1), n. 15. https://doi.org/10.1186/s12910-015-0010-3.
DAVISON, NEIL (2009), *"Non-lethal" Weapons*. Palgrave Macmillan, London.
DAVISON, NEIL (2018), "A legal perspective: Autonomous weapon systems under international humanitarian law". UNODA Occasional Papers, n. 30. United Nations, New York.
"DEEPLOCKER: HOW AI CAN POWER A STEALTHY NEW BREED OF MALWARE" (2018), *Security Intelligence* (blog), 8 agosto. https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.
DEFENSE INNOVATION BOARD (2017), "Defense innovation board recommendations". https://media.defense.gov/2017/Dec/18/2001857962/-1/-1/0/2017-2566-148525RECOMMENDATION%2012_(2017-09-19-01-45-51).PDF.
DEFENSE INNOVATION BOARD (2019), "AI Principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense".

https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. Ultimo accesso 9 giugno 2024.
DEFENSE TECHNICAL INFORMATION CENTER (2013), "Joint publication 2-0 – Joint intelligence". Department of Defense. https://web.archive.org/web/20160613010839/http://www.dtic.mil/doctrine/new_pubs/jp2_0.pdf.
DEHOUSSE, RENAUD (1998), *The European Court of Justice: The Politics of Judicial Integration*. St. Martin's Press, New York.
DEPARTMENT FOR DIGITAL, CULTURE, MEDIA & SPORT (2018), "Data ethics framework". https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.
DEPARTMENT OF NATIONAL DEFENCE (2018), "Autonomous systems for defence and security: Trust and barriers to adoption. Innovation network opportunities". Government of Canada, 16 luglio. https://www.canada.ca/en/department-national-defence/programs/defence-ideas/current-opportunities/innovation-network-opportunities.html#ftn1.
DILLER, ANTONI (1994), *Z: An Introduction to Formal Methods*. 2ª ed. Wiley & Sons, New York.
DING, WEN, SONWOO KIM, DANIEL XU, INKI KIM (2019), "Can intelligent agent improve human-machine team performance under cyberattacks?". In WALDEMAR KARWOWSKI, TAREQ AHRAM (a cura di), *2019 Intelligent Human Systems Integration*. Springer, Amsterdam, pp. 725-730. https://doi.org/10.1007/978-3-030-11051-2_110.
DOCHERTY, BONNIE (2014), "Shaking the foundations: The human rights implications of killer robots". Human Rights Watch. https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots.
DOCHERTY, BONNIE (2020), "The need for and elements of a new treaty on fully autonomous weapons". Human Rights Watch. 1º giugno. https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons.
DOD RESPONSIBLE AI WORKING COUNCIL (2022), "Responsible artificial intelligence strategy and implementation pathway". Carnegie Mellon University, United States, Pittsburgh, PA.
DONALDSON, THOMAS, LEE E. PRESTON (1995), "The stakeholder theory of the corporation: Concepts, evidence, and implications". In *Academy of Management Review*, 20 (1), pp. 65-91.
DOYLE, ANDY, GRAHAM KATZ, KRISTEN SUMMERS, CHRIS ACKERMANN, ILYA ZAVORIN, ZUNSIK LIM, SATHAPPAN MUTHIAH, ET AL. (2014), "Forecasting significant societal events using the embers streaming predictive analytics system". In *Big Data*, 2 (4), pp. 185-195. https://doi.org/10.1089/big.2014.0046.
DUNNMON, JARED, BRYCE GOODMAN, PETER KIRECHU, CAROL SMITH, ALEXANDREA VAN DEUSEN (2021), "Responsible AI guidelines in practice: Operationalizing DoD's ethical principles for AI". Defense Innovation Unit. https://assets.ctfassets.net/3nanhbfkr0pc/acoo1Fj5uungnGNPJ3QWy/3a1dafd64f22efcf8f27380aafae9789/2021_RAI_Report-v3.pdf.
EHSAN, UPOL, MARK O. RIEDL (2020), "Human-centered explainable AI: Towards a reflective sociotechnical approach". In CONSTANTINE STEPHANIDIS, MASAAKI KUROSU, HELMUT DEGEN, LAUREN REINERMAN-JONES (a cura di), *HCI International 2020 – Late Breaking Papers: Multimodality and Intelligence*. Springer International Publishing, Cham, pp. 449-466. https://doi.org/10.1007/978-3-030-60117-1_33.
EITEL-PORTER, RAY (2021), "Beyond the promise: Implementing ethical AI". In *AI and Ethics*, 1 (1), pp. 73-80. https://doi.org/10.1007/s43681-020-00011-6.
EKELHOF, MEREL (2019), "Moving beyond semantics on autonomous weapons: Meaningful human control in operation". In *Global Policy*, 10 (3), pp. 343-348.

https://doi.org/10.1111/1758-5899.12665.
EKELHOF, MEREL, GIACOMO PERSI PAOLI (2020), "The human element in decisions about the use of force". 31 marzo. UNIDIR. https://unidir.org/publication/the-human-element-in-decisions-about-the-use-of-force/.
ENEMARK, CHRISTIAN (2008), "'Non-lethal' weapons and the occupation of Iraq: Technology, ethics and law". In *Cambridge Review of International Affairs*, 21 (2), pp. 199-215.
ENEMARK, CHRISTIAN (2011), "Drones over Pakistan: Secrecy, ethics, and counterinsurgency". In *Asian Security*, 7 (3), pp. 218-237. https://doi.org/10.1080/14799855.2011.615082.
ENISA (2020), "Artificial intelligence cybersecurity challenges". https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.
ERWIN, SANDRA (2017), "With commercial satellite imagery, computer learns to quickly find missile sites in China". In *SpaceNews*, 19 ottobre. https://spacenews.com/with-commercial-satellite-imagery-computer-learns-to-quickly-find-missile-sites-in-china/.
EUROPEAN COMMISSION (2021), *AI Act Proposal*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.
EUROPEAN UNION (2014), "Cyber defence in the EU: Preparing for cyber warfare? Think tank". Brussels. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2014)542143.
EUROPEAN UNION (2015), "Cyber diplomacy: EU dialogue with third countries – Think tank". Brussels. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2015)564374.
EVANS, MICHAEL (2021), "Pentagon uses AI to predict enemy moves 'days in advance'". In *The Times* (London), sez. *World*, 3 agosto. https://www.thetimes.co.uk/article/pentagon-uses-ai-to-predict-enemy-moves-days-in-advance-bql5q5s9p.
EYKHOLT, KEVIN, IVAN EVTIMOV, EARLENCE FERNANDES, BO LI, AMIR RAHMATI, CHAOWEI XIAO, ATUL PRAKASH, TADAYOSHI KOHNO, DAWN SONG (2018), "Robust physical-world attacks on deep learning visual classification". In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, pp. 1625-1634. https://doi.org/10.1109/CVPR.2018.00175.
FABRE, CÉCILE (2009), "Guns, food, and liability to attack in war". In *Ethics*, 120 (1), pp. 36-63. https://doi.org/10.1086/649218.
FANG, RICHARD, ROHAN BINDU, AKUL GUPTA, QIUSI ZHAN, DANIEL KANG (2024), "Teams of LLM agents can exploit zero-day vulnerabilities". In *arXiv*. https://doi.org/10.48550/ARXIV.2406.01637.
FAZELPOUR, SINA, ZACHARY C. LIPTON (2020), "Algorithmic fairness from a non-ideal perspective". In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 57-63. Association for Computing Machinery, New York. https://doi.org/10.1145/3375627.3375828.
FEDERAL FOREIGN OFFICE (2020), "German commentary on operationalizing all eleven guiding principles at a national level as requested by the chair of the 2020 Group of Governmental Experts (GGE) on emerging technologies in the area of Lethal Autonomous Weapons Systems (LAWS) within the Convention on Certain Conventional Weapons (CCW)". https://documents.unoda.org/wp-content/uploads/2020/07/20200626-Germany.pdf.
FEDERAZIONE RUSSA (2017), "Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the convention. Submitted by the Russian Federation". In *Item 6. Examination of Various Dimensions of Emerging Technologies in the Area of Lethal Autonomous Weapons*

Systems, in the Context of the Objective and Purposes of the Convention. Genève. https://admin.govexec.com/media/russia.pdf.
FENNELLY, NIAL (1997), "Legal interpretation at the European Court of Justice". In *Fordham International Law Journal*, 20 (3), pp. 656-679.
FISCHER, JOHN MARTIN, MARK RAVIZZA (2000), *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, Cambridge.
FLORIDI, LUCIANO (2006), "Information ethics, its nature and scope". In *SIGCAS Computers and Society*, 36 (3), pp. 21-36. https://doi.org/10.1145/1195716.1195719.
FLORIDI, LUCIANO (2008), "The method of Levels of Abstraction". In *Minds and Machines*, 18 (3), pp. 303-329. https://doi.org/10.1007/s11023-008-9113-7.
FLORIDI, LUCIANO (2012), "Distributed morality in an information society". In *Science and Engineering Ethics*, 19 (3), pp. 727-743. https://doi.org/10.1007/s11948-012-9413-4.
FLORIDI, LUCIANO (2013), *The Ethics of Information*. Oxford University Press, Oxford.
FLORIDI, LUCIANO (2014), *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press, Oxford (*La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*. Tr. it. Raffaello Cortina, Milano 2017).
FLORIDI, LUCIANO (2016), "Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions". In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2083), 20160112. https://doi.org/10.1098/rsta.2016.0112.
FLORIDI, LUCIANO (2017), "Infraethics – on the conditions of possibility of morality". In *Philosophy & Technology*, 30 (4), pp. 391-394. https://doi.org/10.1007/s13347-017-0291-1.
FLORIDI, LUCIANO (2019), "Translating principles into practices of digital ethics: Five risks of being unethical". In *Philosophy & Technology*, 32 (2), pp. 185-193. https://doi.org/10.1007/s13347-019-00354-x.
FLORIDI, LUCIANO, JOSH COWLS (2019), "A unified framework of five principles for AI in society". In *Harvard Data Science Review*, giugno. https://doi.org/10.1162/99608f92.8cd550d1.
FLORIDI, LUCIANO, JOSH COWLS, THOMAS C. KING, MARIAROSARIA TADDEO (2020), "How to design AI for social good: Seven essential factors". In *Science and Engineering Ethics*, 26 (3), pp. 1771-1796. https://doi.org/10.1007/s11948-020-00213-5.
FLORIDI, LUCIANO, MATTHIAS HOLWEG, MARIAROSARIA TADDEO, JAVIER AMAYA SILVA, JAKOB MÖKANDER, YUNI WEN (2022), "CapAI – a procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act". 23 marzo. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091.
FLORIDI, LUCIANO, JEFF W. SANDERS (2004), "On the morality of artificial agents". In *Minds and Machines*, 14 (3), pp. 349-379.
FLORIDI, LUCIANO, MARIAROSARIA TADDEO (2014) (a cura di), *The Ethics of Information Warfare*. Springer, New York.
FLORIDI, LUCIANO, MARIAROSARIA TADDEO (2016), "What is data ethics?". In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2083), 20160360. https://doi.org/10.1098/rsta.2016.0360.
FLORIDI, LUCIANO, MARIAROSARIA TADDEO (2018), "Romans would have denied robots legal personhood". In *Nature*, 557 (7705), p. 309. https://doi.org/10.1038/d41586-018-05154-5.
FOREIGN & COMMONWEALTH OFFICE (2016), "United Kingdom of Great Britain and Northern Ireland Statement to the informal meeting of experts on lethal autonomous weapons systems, 11-15 April 2016". https://unog.ch/80256EDD006B8954/(httpAssets)/44E4700A0A8CED0EC1257F940053FE3B/$file/2016_LAWS+MX_Towardaworkingdefinition_StatementsUnited+Kindgom.pdf.

FOY, JAMES (2014), “Autonomous weapons systems: Taking the human out of international humanitarian law”. In *Dalhousie Journal of Legal Studies*, 23, pp. 47-70.

FRA (European Union Agency for Fundamental Rights) (2019), “Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights”. In *FRA Focus*. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf.

FRAGA-LAMAS, PAULA, TIAGO M. FERNÁNDEZ-CARAMÉS, MANUEL SUÁREZ-ALBELA, LUIS CASTEDO, MIGUEL GONZÁLEZ-LÓPEZ (2016), “A review on internet of things for defense and public safety”. In *Sensors* (Basel), 16 (10). https://doi.org/10.3390/s16101644.

FREEDBERG, SYDENEY (2014), “NATO hews to strategic ambiguity on cyber deterrence”. https://breakingdefense.com/2014/11/natos-hews-to-strategic-ambiguity-on-cyber-deterrence/. Ultimo accesso luglio 2024.

FREEDMAN, LAWRENCE (2004), *Deterrence*. Polity Press, Cambridge.

FREEMAN, LINDSAY (2021), “Weapons of war, tools of justice: Using artificial intelligence to investigate international crimes”. In *Journal of International Criminal Justice*, 19 (1), pp. 35-53. https://doi.org/10.1093/jicj/mqab013.

G7 DECLARATION (2017), “G7 declaration on responsible state behavior in cyberspace”. Lucca. http://www.mofa.go.jp/files/000246367.pdf.

GALLIOTT, JAI (2017), *Military Robots: Mapping the Moral Landscape*. http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781317096009.

GARLAND, DAVID (2003), “The rise of risk”. In RICHARD V. ERICSON, AARON DOYLE (a cura di), *Risk and Morality*. University of Toronto Press, Toronto, pp. 48-86.

GAVAGHAN, COLIN, ALISTAIR KNOTT, JAMES MACLAURIN, JOHN ZERILLI, JOY LIDDICOAT (2019), “Government use of artificial intelligence in New Zealand”. Final Report on Phase 1 of the Law Foundation’s Artificial Intelligence and Law in New Zealand Project. Law Foundation, Wellington, New Zealand. https://www.cs.otago.ac.nz/research/ai/AI-Law/NZLF%20report.pdf.

GCHQ (2021), “Pioneering a new national security: The ethics of artificial intelligence”. In *GCHQ*. https://www.gchq.gov.uk/files/GCHQAIPaper.pdf.

GEERS, K. (2012), *Sun Tzu and Cyber War*. Cooperative Cyber Defence Centre of Excellence, Tallinn.

GEORGIEVA, ILINA, CLAUDIO LAZO, TJERK TIMAN, ANNE FLEUR VAN VEENSTRA (2022), “From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience”. In *AI and Ethics*, 2 (4), pp. 697-711. https://doi.org/10.1007/s43681-021-00127-3.

GLAESSGEN, EDWARD, DAVID STARGEL (2012), “The digital twin paradigm for future NASA and U.S. air force vehicles”. In *53$_{rd}$ AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. American Institute of Aeronautics and Astronautics, Honolulu, HI. https://doi.org/10.2514/6.2012-1818.

GLERUP, CECILIE, MAJA HORST (2014), “Mapping ‘social responsibility’ in science”. In *Journal of Responsible Innovation*, 1 (1), pp. 31-50. https://doi.org/10.1080/23299460.2014.882077.

GODDARD, KATE, ABDUL ROUDSARI, JEREMY C. WYATT (2012), “Automation bias: A systematic review of frequency, effect mediators, and mitigators”. In *Journal of the American Medical Informatics Association*, 19 (1), pp. 121-127.

GOMEZ, STEVEN R., VINCENT MANCUSO, DIANE STAHELI (2019), “Considerations for human-machine teaming in cybersecurity”. In DYLAN D. SCHMORROW, CALI M. FIDOPIASTIS (a cura di), *Augmented Cognition*. Springer, Cham, 11580, pp. 153-168. https://doi.org/10.1007/978-3-030-22419-6_12.

GOODMAN, WILL (2010), “Will Goodman, cyber deterrence: Tougher in theory than in practice?”. In *Strategic Studies Quarterly*, autunno, pp. 102-135.
GRUT, CHANTAL (2013), “The challenge of autonomous lethal robotics to international humanitarian law”. In *Journal of Conflict and Security Law*, 18 (1), pp. 5-23.
GUASTINI, ROBERTO (2019), “Identificazione, interpretazione dei principi costituzionali”. Università degli Studi di Roma 3, Roma.
GUO, WEISI, KRISTIAN GLEDITSCH, ALAN WILSON (2018), “Retool AI to forecast and limit wars”. In *Nature*, 15 ottobre. https://www.nature.com/articles/d41586-018-07026-4.
GUTHRIE, CHARLES, MICHAEL QUINLAN (2007), *Just War: The Just War Tradition. Ethics in Modern Warfare*. Bloomsbury, London.
HABERMAS, JÜRGEN (1990), “Discourse ethics: Notes on a program of philosophical justification”. In *Moral Consciousness and Communicative Action*. Tr. ing. di C. Lenhardt e S.W. Nicholsen. MIT Press, Cambridge, MA, pp. 43-115.
HABERMAS, JÜRGEN (1998), *The Inclusion of the Other: Studies in Political Theory*. A cura di Ciaran Cronin e Pablo De Greiff. Princeton University Press, Princeton, NJ (*L’inclusione dell’altro. Studi di teoria politica*. Tr. it. Feltrinelli, Milano 2013).
HABERMAS, JÜRGEN (2021), *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Tr. ing. di Thomas Burger e Frederick G. Lawrence. Polity Press, Cambridge (*Storia e critica dell’opinione pubblica*. Tr. it. Laterza, Roma-Bari 2000).
HADDON, CATHERINE (2020), “Ministerial accountability”. Institute for Government. 16 settembre. https://www.instituteforgovernment.org.uk/explainers/ministerial-accountability.
HADFIELD-MENELL, DYLAN, SMITHA MILLI, PIETER ABBEEL, STUART RUSSELL, ANCA DRAGAN (2020), “Inverse reward design”. In *arXiv:1711.02827 [Cs]*, ottobre. http://arxiv.org/abs/1711.02827.
HAGGARD, SIMON, BETH. A. SIMMONS (1987), “Theories of international regimes”. In *International Organization*, 41 (03), p. 491.
HALA SYSTEMS (2022), “Hala systems”. https://halasystems.com/.
HALEY, CRISTOPHER (2013), “A theory of cyber deterrence”. In *Georgetown Journal of International Affairs*, febbraio. http://journal.georgetown.edu/a-theory-of-cyber-deterrence-christopher-haley/.
HARKNETT, RICHARD, J. (1996), “Information warfare and deterrence”. In *U.S. Army War College Parameters*, 10 (26), pp. 93-107.
HARKNETT, RICHARD J., EMILY O. GOLDMAN (2016), “The search for cyber fundamentals”. In *Journal of Information Warfare*, 15 (2), pp. 81-88.
HARWELL, DREW, EVA DOU (2020), “Huawei tested AI software that could recognize Uighur minorities and alert police, report says”. In *The Washington Post*. https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/.
HATHAWAY, OONA, REBECCA CROOTOF (2012), “The law of cyber-attack”. In *California Law Review*, 100 (1), pp. 817-886.
HATHAWAY, OONA A., SCOTT J. SHAPIRO (2018), *The Internationalists: How a Radical Plan to Outlaw War Remade the World*. Simon&Schuster New York.
HEATH, DAVID, DEREK ALLUM, LYNNE DUNCKLEY (1994), *Introductory Logic and Formal Methods*. Alfred Waller, Henley-on-Thames.
HEATH, JOSEPH (2014), “Rebooting discourse ethics”. In *Philosophy & Social Criticism*, 40 (9), pp. 829-866. https://doi.org/10.1177/01914 53714545340.
HEAVEN, WILL DOUGLAS (2021), “DeepMind says its new language model can beat others 25 times its size”. In *MIT Technology Review*, 8 dicembre.

https://www.technologyreview.com/2021/12/08/1041557/deepmind-language-model-beat-others-25-times-size-gpt-3-megatron/.
HEPENSTAL, SAM, LEISHI ZHANG, NEESHA KODAGODA, B.L. WILLIAM WONG (2020), "Pan: Conversational agent for criminal investigations". In FABIO PARTENÒ (a cura di), *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, pp. 134-135. Association for Computing Machinery, New York.
HERSH, SEYMOUR (2000), "Overwhelming force: What happened in the final days of the Gulf War?". In *The New Yorker*, 22 maggio.
HEYNS, CHRISTOF (2014), "Autonomous weapons systems and human rights law". Presentation made at the Informal Expert Meeting Organized by the State Parties to the Convention on Certain Conventional Weapons, 13-16 maggio, Genève.
HEYNS, CHRISTOF (2016a), "Autonomous weapons systems: Living a dignified life and dying a dignified death". In NEHAL BHUTA, SUSANNE BECK, ROBIN GEIß, HIN-YAN LIU, CLAUS KREß (a cura di), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, Cambridge, pp. 3-19.
HEYNS, CHRISTOF (2016b), "Human rights and the use of Autonomous Weapons Systems (AWS) during domestic law enforcement". In *Human Rights Quarterly*, 38 (2), pp. 350-378. https://doi.org/10.1353/hrq.2016.0034.
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (2019), *Ethics Guidelines for Trustworthy AI*. European Commission, Brussels. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.
HM GOVERNMENT (2022), "National cyber strategy". 130. https://www.gov.uk/government/publications/national-cyber-strategy-2022. Ultimo accesso luglio 2024.
HOARE, C.A.R. (1972), "Notes on data structuring". In O.-J. DAHL, E.W. DIJKSTRA, C.A.R. HOARE (a cura di), *Structured Programming*. Academic Press, London, pp. 83-174. http://dl.acm.org/citation.cfm?id=1243380.1243382.
HOFFMAN, WYATT (2021), "AI and the future of cyber competition". Center for Security and Emerging Technology. https://doi.org/10.51593/2020CA007.
HOLLAND MICHEL, ARTHUR (2020a), "The black box, unlocked". UNIDIR. https://unidir.org/publication/black-box-unlocked.
HOLLAND MICHEL, ARTHUR (2020b), "The black box, unlocked: Predictability and understandability in military AI". United Nations Institute for Disarmament Research. https://doi.org/10.37559/SecTec/20/AI1.
HOLLIS, DUNCAN B. (2011), "An E-SOS for cyberspace". In *Harvard International Law Journal*, 52 (373), pp. 374-375.
HOUSE OF LORDS (2019), "Autonomous weapons: Questions for Ministry of Defence UIN HL15333". 24 aprile. https://questions-statements.parliament.uk/written-questions/detail/2019-04-24/HL15333.
HÜLLERMEIER, EYKE, WILLEM WAEGEMAN (2021), "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In *Machine Learning*, 110 (3), pp. 457-506. https://doi.org/10.1007/s10994-021-05946-3.
HUME, DAVID (2009), *A Treatise of Human Nature*. A cura di David Fate Norton. Oxford University Press, Oxford (*Trattato sulla natura umana*. Tr. it. Bompiani, Milano 2005).
HURKA, THOMAS (2005), "Proportionality in the morality of war". In *Philosophy & Public Affairs*, 33 (1), pp. 34-66.
HURKA, THOMAS (2008), "Proportionality and necessity". In LARRY MAY, EMILY CROOKSTON (a cura di), *War: Essays in Political Philosophy*. Cambridge University Press, Cambridge, pp. 127-144.

IASIELLO, EMILIO (2014), “Is cyber deterrence an illusory course of action?”. In *Journal of Strategic Security*, 7 (1), pp. 54-67.
IBM (2021), “What is data labeling?”. 12 agosto. https://www.ibm.com/cloud/learn/data-labeling.
INDEPENDENT SURVEILLANCE REVIEW (2015), “A democratic licence to operate: Report of the Independent Surveillance Review”. Royal United Services Institute for Defence Studies, London. https://static.rusi.org/20150714_whr_2-15_a_democratic_licence_to_operate.pdf.
INSIKT GROUP (2022), “HermeticWiper and PartyTicket targeting computers in Ukraine”. 2 marzo. https://go.recordedfuture.com/hubfs/reports/mtp-2022-0302.pdf.
INTERNATIONAL COMMITTEE OF THE RED CROSS (2016), “Views of the ICRC on autonomous weapon systems”. Aprile. https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system.
INTERNATIONAL COMMITTEE OF THE RED CROSS (2018), “Ethics and autonomous weapon systems: An ethical basis for human control?”. International Committee of the Red Cross, Genève.
INTERNATIONAL COMMITTEE OF THE RED CROSS (2019), “Autonomy, artificial intelligence and robotics: Technical aspects of human control”. https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.
INTERNATIONAL COMMITTEE OF THE RED CROSS (2020), “Treaties, states parties, and commentaries – St Petersburg Declaration relating to explosive projectiles, 1868 – Declaration”. International Committee of the Red Cross, Genève. https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Article.xsp?action=openDocument&documentId=568842C2B90F4A29C12563CD0051547C. Ultimo accesso 2 dicembre 2022.
INTERNATIONAL COMMITTEE OF THE RED CROSS (2021), “ICRC position on autonomous weapon systems & background paper”. International Committee of the Red Cross, Genève.
INTERNATIONAL MILITARY TRIBUNAL (Nuremberg) (1947), “Judgment and sentences, October 1, 1946”. In *American Journal of International Law*, 41, pp. 172-306.
INTERNATIONAL SECURITY ADVISORY BOARD (2014), “A framework for international cyber stability”. US Department of State. http://goo.gl/azdM0B.
ISH, DANIEL, JARED ETTINGER, CHRISTOPHER FERRIS (2021), “Evaluating the effectiveness of artificial intelligence systems in intelligence analysis”. Rand Corporation. https://www.rand.org/pubs/research_reports/RRA464-1.html.
JACKY, JONATHAN (1997), *The Way of Z: Practical Programming with Formal Methods*. Cambridge University Press, Cambridge.
JAGIELSKI, MATTHEW, ALINA OPREA, BATTISTA BIGGIO, CHANG LIU, CRISTINA NITA-ROTARU, BO LI (2018), “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning”. In *arXiv:1804.00308 [Cs]*, aprile. http://arxiv.org/abs/1804.00308.
JAIN, NEHA (2016), “Autonomous weapons systems: New frameworks for individual responsibility”. In NEHAL BHUTA, SUSANNE BECK, ROBIN GEIß, HIN-YAN LIU, CLAUS KREß (a cura di), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, Cambridge, pp. 303-324.
JAPANESE SOCIETY FOR ARTIFICIAL INTELLIGENCE (2017), “Ethical guidelines”. http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf.
JENSEN, ERIC TALBOT (2009), “Cyber warfare and precautions against the effects of attacks”. In *Texas Law Review*, 88 (1533), pp. 1534-1569.
JENSEN, ERIC TALBOT (2012), “Cyber deterrence”. In *Emory International Law Review*, 26 (2), pp. 773-824.

JERVIS, ROBERT (1979), “Deterrence theory revisited”. In *World Politics*, 31 (2), pp. 289-324. https://doi.org/10.2307/2009945.
JERVIS, ROBERT (1988), “Realism, game theory, and cooperation”. In *World Politics*, 40 (3), pp. 317-349. https://doi.org/10.2307/2010216.
JIA, YIFAN, CHRISTOPHER M. POSKITT, JUN SUN, SUDIPTA CHATTOPADHYAY (2022), “Physical adversarial attack on a robotic arm”. In *IEEE Robotics and Automation Letters*, 7 (4), pp. 9334-9341.
JOHNSON, AARON M., SIDNEY AXINN (2013), “The morality of autonomous robots”. In *Journal of Military Ethics*, 12 (2), pp. 129-141. https://doi.org/10.1080/15027570.2013.818399.
JOHNSTON, ROB (2005), *Analytic Culture in the US Intelligence Community: An Ethnographic Study*. Center for the Study of Intelligence, Central Intelligence Agency, Washington, DC. https://web.archive.org/web/20070613143919/https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/analytic-culture-in-the-u-s-intelligence-community/chapter_1.htm.
JUSTICE AND HOME AFFAIRS COMMITTEE (2022), “Technology rules? The advent of new technologies in the justice system”. HLPaper180. House of Lords, Westminster.
KAMM, F.M. (2004), “Failures of Just War Theory: Terror, harm, and justice”. In *Ethics*, 114 (4), pp. 650-692. https://doi.org/10.1086/383441.
KANIA, ELSA B. (2018a), “China’s embrace of AI: Enthusiasm and challenges – European Council on Foreign Relations”. In *ECFR* (blog), 6 novembre. https://ecfr.eu/article/commentary_chinas_embrace_of_ai_enthusiasmandchallenges/.
KANIA, ELSA B. (2018b), “China’s strategic ambiguity and shifting approach to lethal autonomous weapons systems”. In *Lawfare* (blog), 17 aprile. https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.
KANT, IMMANUEL (2019), *Grundlegung zur Metaphysik der Sitten (Großdruck)*. A cura di Theodor Borken. Henricus, Berlin (*Fondazione della metafisica dei costumi*. Tr. it. Bompiani, Milano 2003).
KASHER, ASA (2007), “The principle of distinction”. In *Journal of Military Ethics*, 6 (2), pp. 152-167.
KASTENBERG, JOSHUA E. (2009), “Changing the paradigm of internet access from government information systems: A solution to the need for the DoD to take time-sensitive action on the Niprnet”. In *Air Force Law Review*, 64, pp. 175-210.
KAURIN, PAULINE (2010), “With fear and trembling: An ethical framework for non-lethal weapons”. In *Journal of Military Ethics*, 9 (1), pp. 100-114. https://doi.org/10.1080/15027570903523057.
KAURIN, PAULINE (2015), “And next please: The future of the NLW debate international regulation of emerging military technologies”. In *Case Western Reserve Journal of International Law*, 47 (1), pp. 217-228.
KELION, LEO (2021), “Huawei patent mentions use of Uighur-spotting tech”. In *BBC News*, sez. *Technology*, 13 gennaio. https://www.bbc.com/news/technology-55634388.
KELLY, ERIN I. (2012), “What is an excuse?”. In D. JUSTIN COATES, NEAL A. TOGNAZZINI (a cura di), *Blame*. Oxford University Press, New York, pp. 244-262. https://doi.org/10.1093/acprof:oso/9780199860821.003.0013.
KELLY, JONATHAN, MICHAEL DELAUS, ERIK HEMBERG, UNA-MAY O’REILLY (2019), “Adversarially adapting deceptive views and reconnaissance scans on a software defined network”. In NUR ZINCIR-HEYWOOD, IDILIO DRAGO, ROBERT HARPER (a cura di), *FIP/IEEE Symposium on Integrated Network and Service Management (IM)*. IEEE, Piscataway, NJ, pp. 49-54.

KHOSROW-POUR, MEHDI, D.B.A. (2021) (a cura di), *Encyclopedia of Information Science and Technology*. 5ª ed. IGI Global, Hersey, PA. https://doi.org/10.4018/978-1-7998-3479-3.
KHOURY, ANDREW C. (2018), "The objects of moral responsibility". In *Philosophical Studies*, 175 (6), pp. 1357-1381. https://doi.org/10.1007/s11098-017-0914-5.
KIM, SCOTT Y.H., IAN F. WALL, AIMEE STANCZYK, RAYMOND DE VRIES (2009), "Assessing the public's views in research ethics controversies: Deliberative democracy and bioethics as natural allies". In *Journal of Empirical Research on Human Research Ethics*, 4 (4), pp. 3-16. https://doi.org/10.1525/jer.2009.4.4.3.
KING, TARIQ M., JASON ARBON, DIONNY SANTIAGO, DAVID ADAMO, WENDY CHIN, RAM SHANMUGAM (2019), "AI for testing today and tomorrow: Industry perspectives". In *2019 IEEE International Conference on Artificial Intelligence Testing (AITest)*. IEEE, Newark, CA, pp. 81-88. https://doi.org/10.1109/AITest.2019.000-3.
KIRAT, DHILUNG, JIYONG JANG, MARC PH. STOECKLIN (2018), "DeepLocker: Concealing targeted attacks with AI locksmithing". IBM. https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf.
KLAMM, J., C. DOMINGUEZ, B. YOST, P. MCDERMOTT, M. LENOX (2019), "Partnering with technology: The importance of human machine teaming in future MDC2 systems". In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006, pp. 259-266. SPIE. https://doi.org/10.1117/12.2518750.
KNIEP, RONJA (2019), "Another layer of opacity: How spies use AI and why we should talk about it". In *About:Intel* (blog), 20 dicembre. https://aboutintel.eu/how-spies-use-ai/.
KNIGHT, WILL (2022), "Russia's killer drone in Ukraine raises fears about AI in warfare". In *Wired*, 17 marzo. https://www.wired.com/story/ai-drones-russia-ukraine/.
KONAEV, MARGARITA, HUSANJOT CHAHAL (2021), "Building trust in human-machine teams". https://www.brookings.edu/techstream/building-trust-in-human-machine-teams/.
KORPELA, CHRISTOPHER (2017), "Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)". CCW/GGE.1/2017/CRP.1. United Nations Office for Disarmament Affairs, Genève. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_CRP.1_0.pdf.
KOTT, ALEXANDER (2018), "Intelligent autonomous agents are key to cyber defense of the future army networks". In *arXiv:1812.08014 [Cs]*, dicembre. http://arxiv.org/abs/1812.08014.
KOTT, ALEXANDER, PAUL THÉRON, LUIGI V. MANCINI, EDLIRA DUSHKU, AGOSTINO PANICO, MARTIN DRAŠAR, BENOÎT LEBLANC, ET AL. (2020), "An introductory preview of autonomous intelligent cyber-defense agent reference architecture, Release 2.0". In *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 17 (1), pp. 51-54. https://doi.org/10.1177/1548512919886163.
KRISHNAN, ARMIN (2009), *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Ashgate, Burlington, VT.
KRISHNAN, MAYA (2020), "Against interpretability: A critical examination of the interpretability problem in machine learning". In *Philosophy & Technology*, 33 (3), pp. 487-502. https://doi.org/10.1007/s13347-019-00372-9.
KUGLER, RICHARD (2009), "Deterrence of cyber attacks". In FRANKLIN KRAMER, STUART STARR, LARRY WENTZ (a cura di), *Cyberpower and National Security*. National Defense University, Washington, DC, pp. 309-342.
LA FORS, KAROLINA, BART CUSTERS, ESTHER KEYMOLEN (2019), "Reassessing values for emerging big data technologies: Integrating design-based and application-based approaches". In *Ethics and Information Technology*, 21 (3), pp. 209-226. https://doi.org/10.1007/s10676-019-09503-4.

LAIRD, JOHN, CHARAN RANGANATH, SAMUEL GERSHMAN (2019), “Future directions in human machine teaming workshop”. https://basicresearch.defense.gov/Portals/61/Future%20Directions%20in%20Human%20Machine%20Teaming%20Workshop%20report%20%20%28for%20public%20release%29.pdf.

LAN, TANG, ZHANG XIN, HARRY RADUEGE JR., DMITRY GRIGORIEV, PAVAN DUGGAL, STEIN SCHJØLBERG (2010), *Global Cyber Deterrence Views from China, the U.S., Russia, India, and Norway*. EastWest Institute, New York.

LANGO, JOHN W. (2010), “Nonlethal weapons, noncombatant immunity, and combatant nonimmunity: A study of Just War Theory”. In *Philosophia*, 38 (3), pp. 475-497. https://doi.org/10.1007/s11406-009-9231-3.

LAVIN, ALEXANDER, HECTOR ZENIL, BROOKS PAIGE, DAVID KRAKAUER, JUSTIN GOTTSCHLICH, TIM MATTSON, ANIMA ANANDKUMAR, ET AL. (2021), “Simulation intelligence: Towards a new generation of scientific methods”. In *arXiv:2112.03235 [Cs]*, dicembre. http://arxiv.org/abs/2112.03235.

LEBRETON, GILLES (2021), “Report of the Committee on Legal Affairs to the European Parliament”. https://www.europarl.europa.eu/doceo/document/A-9-2021-0001EN.html#.

LEVY, NEIL (2008), “The responsibility of the psychopath revisited”. In *Philosophy, Psychiatry, & Psychology*, 14 (2), pp. 129-138. https://doi.org/10.1353/ppp.0.0003.

LIAO, CONG, HAOTI ZHONG, ANNA SQUICCIARINI, SENCUN ZHU, DAVID MILLER (2018), “Backdoor embedding in convolutional neural network models via invisible perturbation”. In *arXiv:1808.10307 [Cs, Stat]*, agosto. http://arxiv.org/abs/1808.10307.

LIBICKI, MARTIN C. (1997), “Defending cyberspace and other metaphors”. Institute for National Strategic Studies, National Defense University, Washington, DC.

LIBICKI, MARTIN C. (2009), *Cyberdeterrence and Cyberwar*. Rand, Santa Monica, CA. http://www.rand.org/pubs/monographs/MG877.html.

LIEBLICH, ELIAV, EYAL BENVENISTI (2016), “The obligation to exercise discretion in warfare: Why autonomous weapons systems are unlawful”. In NEHAL BHUTA, SUSANNE BECK, ROBIN GEIß, HIN-YAN LIU, CLAUS KREß (a cura di), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, Cambridge, pp. 245-283.

LIN, HERBERT (2012), “Cyber conflict and international humanitarian law”. In *International Review of the Red Cross*, 94 (886), pp. 515-531. https://doi.org/10.1017/S1816383112000811.

LIPPERT-RASMUSSEN, KASPER (2014), “Just War Theory, intentions, and the deliberative perspective objection”. In HELEN FROWE, GERALD LANG (a cura di), *How We Fight: Ethics in War*. Oxford University Press, Oxford, pp. 138-154.

LIST, CHRISTIAN, PHILIP PETTIT (2011), *Group Agency*. Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780199591565.001.0001.

LIU, HIN-YAN (2016), “Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems”. In NEHAL BHUTA, SUSANNE BECK, ROBIN GEIß, HIN-YAN LIU, CLAUS KREß (a cura di), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, Cambridge, pp. 325-344.

LLORENS, ALBERTINA ALBORS (1999), “The European Court of Justice, more than a teleological court”. In *Cambridge Yearbook of European Legal Studies*, 2, pp. 373-398. https://doi.org/10.5235/152 888712802815789.

LO, CHRIS (2015), “Safer with data: Protecting Pakistan’s schools with predictive analytics”. In *Army Technology*, 8 novembre. https://www.army-technology.com/features/featuresafer-with-data-protecting-pakistans-schools-with-predictive-analytics-4713601/.

LOPEZ, TODD (2022), “Simplified human/machine interfaces top list of critical DoD technologies”. https://www.defense.gov/News/News-

Stories/Article/Article/2904627/simplified-humanmachine-interfaces-top-list-of-critical-dod-technologies/.
"LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS" (2012). Human Rights Watch. 19 novembre. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.
LYSAGHT, ROBERT J., REGINA HARRIS, WILLIAM KELLY (1988), "Artificial intelligence for command and control". Analytics, Willow Grove, PA. https://apps.dtic.mil/docs/citations/ADA229342.
MAKARIUS, ERIN E., DEBMALYA MUKHERJEE, JOSEPH D. FOX, ALEXA K. FOX (2020), "Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization". In *Journal of Business Research*, 120 (novembre), pp. 262-273. https://doi.org/10.1016/j.jbusres.2020.07.045.
MÄNTYMÄKI, MATTI, MATTI MINKKINEN, TEEMU BIRKSTEDT, MIKA VILJANEN (2022), "Defining organizational AI governance". In *AI and Ethics*, 2 (4), pp. 603-609. https://doi.org/10.1007/s43681-022-00143-x.
MARCHANT, GARY E., BRADEN ALLENBY, RONALD ARKIN, EDWARD T. BARRETT (2011), "International governance of autonomous military robots". In *Columbia Science and Technology Law Review*, 12, pp. 272-316.
MARCUM, RICHARD A., CURT H. DAVIS, GRANT J. SCOTT, TYLER W. NIVIN (2017), "Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks". In *Journal of Applied Remote Sensing*, 11 (4), 042614. https://doi.org/10.1117/1.JRS.11.042614.
MARGALIT, AVISHAI, MICHAEL WALZER (2009), "Israel: Civilians & combatants". In *The New York Review of Books*, maggio. https://www.nybooks.com/articles/2009/08/13/israel-civilians-combatants-an-exchange.
MATSUMOTO, MASAKAZU (2020), "Amoral realism or just war morality? Disentangling different conceptions of necessity". In *European Journal of International Relations*, 26 (4), pp. 1084-1105. https://doi.org/10.1177/1354066120910233.
MATTHIAS, ANDREAS (2004), "The responsibility gap: Ascribing responsibility for the actions of learning automata". In *Ethics and Information Technology*, 6 (3), pp. 175-183. https://doi.org/10.1007/s10676-004-3422-1.
MCCARTHY, THOMAS (1995), "Practical discourse: On the relation of morality to politics". In *Revue Internationale de Philosophie*, 49 (194), pp. 461-481.
MCCONNELL, MIKE (2010), "Mike McConnell on how to win the cyber-war we're losing". 28 febbraio. http://www.washingtonpost.com/wp-dyn/content/article/2010/02/25/AR2010022502493.html.
MCINTYRE, ALISON (2004), "Doctrine of double effect". Luglio. https://stanford.library.sydney.edu.au/entries/double-effect/.
MCKENDRICK, KATHLEEN (2019), *Artificial Intelligence Prediction and Counterterrorism*. Chatham House, London. https://www.chathamhouse.org/sites/default/files/2019-08-07-AICounterterrorism.pdf.
MCMAHAN, JEFF (2006), "On the moral equality of combatants". In *Journal of Political Philosophy*, 14 (4), pp. 377-393.
MCMAHAN, JEFF (2009), *Killing in War*. Oxford University Press, Oxford.
MCMAHAN, JEFF (2010), "The just distribution of harm between combatants and noncombatants". In *Philosophy & Public Affairs*, 38 (4), pp. 342-379. https://doi.org/10.1111/j.1088-4963.2010.01196.x.
MCMAHAN, JEFF (2011), "Who is morally liable to be killed in war?". In *Analysis*, 71 (3), pp. 544-559.

MCMAHAN, JEFF, ROBERT MCKIM (1993), "The Just War and the Gulf War". In *Canadian Journal of Philosophy*, 23 (4), pp. 501-541.
MCNEESE, NATHAN J., BEAU G. SCHELBLE, LORENZO BARBERIS CANONICO, MUSTAFA DEMIR (2021), "Who/what is my teammate? Team composition considerations in human-AI teaming". In *arXiv:2105.11000 [Cs]*, maggio. http://arxiv.org/abs/2105.11000.
MEISELS, TAMAR (2018), *Contemporary Just War: Theory and Practice*. Routledge, London.
MILLER, SEUMAS (2018), *Dual Use Science and Technology, Ethics and Weapons of Mass Destruction*. Springer, New York.
MINISTRY OF DEFENCE (2011), "Joint service manual of the law of armed conflict (JSP 383)". https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf.
MINISTRY OF DEFENCE (2018a), "Unmanned aircraft systems (JDP 0-30.2)". https://www.gov.uk/government/publications/unmanned-aircraft-systems-jdp-0-302.
MINISTRY OF DEFENCE (2018b), "Human-machine teaming (JCN 1/18)". https://www.gov.uk/government/publications/human-machine-teaming-jcn-118.
MINISTRY OF DEFENCE (2022), "Ambitious, safe, responsible: Our approach to the delivery of AI-enabled capability in defence". https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082991/20220614-Ambitious_Safe_and_Responsible.pdf.
MIRSKY, YISROEL, TOM MAHLER, ILAN SHELEF, YUVAL ELOVICI (2019), "CT-GAN: Malicious tampering of 3D medical imagery using deep learning". In *ResearchGate*. https://www.researchgate.net/publication/330357848_CT-GAN_Malicious_Tampering_of_3D_Medical_Imagery_using_Deep_Learning.
MITCHELL, KWASI, JOE MARIANI, ADAM ROUTH, AKASH KEYAL, ALEX MIRKOW (2019), *The Future of Intelligence Analysis: A Task-Level View of the Impact of Artificial Intelligence on Intel Analysis*. Deloitte, Washington, DC.
MITCHELL, MARGARET, SIMONE WU, ANDREW ZALDIVAR, PARKER BARNES, LUCY VASSERMAN, BEN HUTCHINSON, ELENA SPITZER, INIOLUWA DEBORAH RAJI, TIMNIT GEBRU (2019), "Model cards for model reporting". In *Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19*, pp. 220-229. https://doi.org/10.1145/3287560.3287596.
MÖKANDER, JAKOB, LUCIANO FLORIDI (2021), "Ethics-based auditing to develop trustworthy AI". In *Minds and Machines*, febbraio. https://doi.org/10.1007/s11023-021-09557-8.
MOOR, JAMES H. (1985), "What is computer ethics?". In *Metaphilosophy*, 16 (4), pp. 266-275. https://doi.org/10.1111/j.1467-9973.1985.tb00173.x.
MOORE, CRISTOPHER (1990), "Unpredictability and undecidability in dynamical systems". In *Physical Review Letters*, 64 (20), pp. 2354-2357. https://doi.org/10.1103/PhysRevLett.64.2354.
MORGAN, PATRICK M. (2003), *Deterrence Now*. Cambridge University Press, Cambridge.
MORGAN, PATRICK M. (2010), "Applicability of traditional deterrence concepts and theory to the cyber realm". In *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for U.S. Policy*. National Academic Press, Washington, DC, pp. 55-76.
MORLEY, JESSICA, JOSH COWLS, MARIAROSARIA TADDEO, LUCIANO FLORIDI (2020), "Ethical guidelines for COVID-19 tracing apps". In *Nature*, 582, pp. 29-31.
MORLEY, JESSICA, ANAT ELHALAL, FRANCESCA GARCIA, LIBBY KINSEY, JAKOB MÖKANDER, LUCIANO FLORIDI (2021), "Ethics as a service: A pragmatic operationalisation of AI ethics". In *Minds and Machines*, 31 (2), pp. 239-256. https://doi.org/10.1007/s11023-021-09563-w.

MORLEY, JESSICA, LUCIANO FLORIDI, LIBBY KINSEY, ANAT ELHALAL (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 (4), pp. 2141-2168. https://doi.org/10.1007/s11948-019-00165-5.

MOSELEY, ALEXANDER (2011), "Just War Theory". In D.J. CHRISTIE (a cura di), *The Encyclopedia of Peace Psychology*. John Wiley & Sons. https://doi.org/10.1002/9780470672532.wbepp144.

MUELLER, JOHN (1995), "The perfect enemy: Assessing the Gulf War". In *Security Studies*, 5 (1), pp. 77-117. https://doi.org/10.1080/09636419508429253.

MUSIOLIK, THOMAS HEINRICH, ADRIAN DAVID CHEOK (2021) (a cura di), *Analyzing Future Applications of AI, Sensors, and Robotics in Society: Advances in Computational Intelligence and Robotics*. IGI Global, Hersey, PA. https://doi.org/10.4018/978-1-7998-3499-1.

NAGEL, THOMAS (1972), "War and massacre". In *Philosophy and Public Affairs*, 1 (inverno), pp. 123-144 ("Guerra e massacro". Tr. it. in *Questioni mortali*. Il Saggiatore, Milano 2015).

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE (2022), *Human-AI Teaming: State-of-the-Art and Research Needs*. Committee on Human-System Integration Research Topics for the 711th Human, Performance Wing of the Air Force Research Laboratory, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, and Board on Human-Systems Integration. National Academies Press, Washington, DC. https://doi.org/10.17226/26355.

NATIONAL SECURITY AGENCY (2012), "(U) SIGINT Strategy. 23 February 2012". In "A strategy for surveillance powers". In *The New York Times*, 23 novembre 2013. http://www.nytimes.com/interactive/2013/11/23/us/politics/23nsa-sigint-strategy-document.html.

NATO (2020), "AAP-06 Edition 2020: NATO Glossary of Terms and Definitions". NATO Standardization Office.

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE (2013), *Tallinn Manual on the International Law Applicable to Cyber Warfare: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence*. Cambridge University Press, Cambridge.

NELKIN, DANA KAY (2011), *Making Sense of Freedom and Responsibility*. Oxford University Press, New York.

NGUYEN, ANH M., JASON YOSINSKI, JEFF CLUNE (2015), "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images". In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA. https://doi.org/10.1109/CVPR.2015.7298640.

NIST (2022), "AI risk management framework: Initial draft". 17 marzo. https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf.

NIU, YARU, ROHAN PALEJA, MATTHEW GOMBOLAY (2021), "Multi-agent graph-attention communication and teaming". In AAMAS, 21.

NOOR, UMARA, ZAHID ANWAR, TEHMINA AMJAD, KIM-KWANG RAYMOND CHOO (2019), "A machine learning-based fintech cyber threat attribution framework using high-level indicators of compromise". In *Future Generation Computer Systems*, 96 (luglio), pp. 227-242. https://doi.org/10.1016/j.future.2019.02.013.

NORVEGIA (2017), "CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems: General statement by Norway". https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_(2017)/2017_GGE%2BLAWS_StatementNorway.pdf. Ultimo accesso luglio 2024.

NURICK, LESTER (1945), “The distinction between combatant and noncombatant in the law of war”. In *American Journal of International Law*, 39 (4), pp. 680-697. https://doi.org/10.2307/2193409.
NYE, JOSEPH S. (2011), “Nuclear lessons for cyber security?”. In *Strategic Studies Quarterly*, 5 (4), pp. 11-38.
O’CONNELL, MARY ELLEN (2012), “Cyber security without cyber war”. In *Journal of Conflict and Security Law*, 17 (2), pp. 187-209. https://doi.org/10.1093/jcsl/krs017.
O’CONNELL, MARY ELLEN (2014), “The American way of bombing: How legal and ethical norms change”. In MATTHEW EVANGELISTA, HENRY SHUE (a cura di), *The American Way of Bombing: Changing Ethical and Legal Norms, from Flying Fortresses to Drones*. Cornell University Press Ithaca, NY, pp. 1-24.
OFFICE OF THE SECRETARY OF DEFENSE (2017), “Department of Defense fiscal year (FY) 2017 request for additional appropriations”.
OHLIN, JENS DAVID, LARRY MAY (2016), *Necessity in International Law*. Oxford University Press, Oxford.
OMAND, DAVID, MARK PHYTHIAN (2018), *Principled Spying: The Ethics of Secret Intelligence*. Oxford University Press, Oxford.
O’NEILL, THOMAS, NATHAN MCNEESE, AMY BARRON, BEAU SCHELBLE (2020), “Human-autonomy teaming: A review and analysis of the empirical literature”. In *Human Factors*, ottobre. https://journals.sagepub.com/doi/full/10.1177/0018720820960865.
OPENAI (2019), “Better language models and their implications”. In *OpenAI* (blog), 14 febbraio. https://openai.com/blog/better-language-models/.
OREND, BRIAN (2001), “Just and lawful conduct in war: Reflections on Michael Walzer”. In *Law and Philosophy*, 20 (1), pp. 1-30. https://doi.org/10.2307/3505049.
OREND, BRIAN (2019), *War and Political Theory*. Polity, Cambridge.
OWENS, WILLIAM A., KENNETH W. DAM, HERBERT LIN (2009) (a cura di), *Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities*. National Academies Press, Washington, DC.
PAESI BASSI (2017), *Examination of Various Dimensions of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, in the Context of the Objectives and Purposes of the Convention*. CCW/GGE.1/2017/WP.2. Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects. United Nations Office for Disarmament Affairs, Genève. https://www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2017/gge/documents/WP2.pdf.
PAYNE, KENNETH (2021), *I, Warbot: The Dawn of Artificially Intelligent Conflict*. Hurst & Company, London.
PELLERIN, CHERYL (2017), “Project Maven industry day pursues artificial intelligence for DoD challenges”. US Department of Defense. https://www.defense.gov/News/News-Stories/Article/Article/1356172/project-maven-industry-day-pursues-artificial-intelligence-for-dod-challenges/.
PERRY, STEPHEN R. (1995), “Risk, harm, and responsibility”. In DAVID G. OWEN (a cura di), *Philosophical Foundations of Tort Law*. Oxford University Press, Oxford, pp. 321-346. https://watermark.silverchair.com/acprof-9780198265795-chapter-15.pdf?t.
PETERS, DORIAN (2019), “Beyond principles: A process for responsible tech”. In *The Ethics of Digital Experience* (blog), 14 maggio. https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317.
POSSONY, STEFAN T. (1946), “Atomic power and world order”. In *Review of Politics*, 8 (4), pp. 533-535.

POWELL, ROBERT (2008), *Nuclear Deterrence Theory: The Search for Credibility*. Cambridge University Press, Cambridge.
RAAIJMAKERS, STEPHAN (2019), "Artificial intelligence for law enforcement: Challenges and opportunities". In *IEEE Security & Privacy*, 17 (5), pp. 74-77.
RAE, JACK, GEOFFREY IRVING, LAURA WEIDINGER (2021), "Language modelling at scale: Gopher, ethical considerations, and retrieval". In *DeepMind* (blog), 8 dicembre. https://deepmind.com/blog/article/language-modelling-at-scale.
RAMSEY, PAUL (2002), *The Just War: Force and Political Responsibility*. Rowman & Littlefield, Lanham, MD.
RASSLER, DON (2021), "Data, AI, and the future of U.S. counterterrorism: Building an action plan". In *CTC Sentinel*, 14 (8), pp. 31-44.
RATTRAY, GREGORY J. (2009), "An environmental approach to understanding cyberpower". In STUART S. KRAMER, LERRY K. WENTZ (a cura di), *Cyberpower and National Security*. National Defense University Press, Washington, DC, pp. 253-274.
RAWLS, JOHN (2005), *A Theory of Justice*. Belknap Press, Cambridge, MA (*Una teoria della giustizia*. Tr. it. Feltrinelli, Milano 2008).
RÉPUBLIQUE FRANÇAISE (2016), "Working paper of France: 'Characterization of a LAWS'". In *Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*. https://unog.ch/80256EDD006B8954/(httpAssets)/5FD844883B46FEACC1257F8F00401FF6/$file/2016LAWSMX_CountryPaperFrance+CharacterizationofaLAWS.pdf.
RICE, H.G. (1956), "On completely recursively enumerable classes and their key arrays". In *Journal of Symbolic Logic*, 21 (3), pp. 304-308. https://doi.org/10.2307/2269105.
RIGAKI, MARIA, AHMED ELRAGAL (2017), "Adversarial deep learning against intrusion detection classifiers". Master's thesis, Luleå University of Technology, Luleå.
ROBBINS, MARTIN (2016), "Has a rampaging AI algorithm really killed thousands in Pakistan?". In *The Guardian*, sez. *Science*, 18 febbraio. https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan.
ROBERTS, HUW, JOSH COWLS, JESSICA MORLEY, MARIAROSARIA TADDEO, VINCENT WANG, LUCIANO FLORIDI (2020), "The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation". In *AI & Society*, giugno. https://doi.org/10.1007/s00146-020-00992-2.
ROBINETTE, PAUL, AYANNA M. HOWARD, ALAN R. WAGNER (2017), "Effect of robot performance on human-robot trust in time-critical situations". In *IEEE Transactions on Human-Machine Systems*, 47 (4), pp. 425-436. https://doi.org/10.1109/THMS.2017.2648849.
ROBINETTE, PAUL, WENCHEN LI, ROBERT ALLEN, AYANNA M. HOWARD, ALAN R. WAGNER (2016), "Overtrust of robots in emergency evacuation scenarios". In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 101-108. IEEE, Christchurch, New Zealand.
ROFF, HEATHER M. (2014), "The strategic robot problem: Lethal autonomous weapons in war". In *Journal of Military Ethics*, 13 (3), pp. 211-227. https://doi.org/10.1080/15027570.2014.975010.
ROFF, HEATHER M. (2015), "Lethal autonomous weapons and *jus ad bellum* proportionality". In *Case Western Reserve Journal of International Law*, 47 (1), pp. 37-52.
ROFF, HEATHER M. (2020a), *Uncomfortable Ground Truths: Predictive Analytics and National Security*. Brookings Institute, Washington, DC.
ROFF, HEATHER M. (2020b), "Forecasting and predictive analytics: A critical look at the basic building blocks of a predictive model". In *Brookings* (blog), 11 settembre. https://www.brookings.edu/techstream/forecasting-and-predictive-analytics-a-critical-look-at-the-basic-building-blocks-of-a-predictive-model/.

ROWLANDS, MARK (2000), *The Environmental Crisis: Understanding the Value of Nature*. Palgrave Macmillan, New York.
RUDIN, CYNTHIA (2019), "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In *Nature Machine Intelligence*, 1 (5), pp. 206-215. https://doi.org/10.1038/s42256-019-0048-x.
RUDIN, CYNTHIA, MIT SLOAN (2013), "Predictive policing: Using machine learning to detect patterns of crime". In *Wired*, 22 agosto. https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/.
RUDIN, CYNTHIA, CAROLINE WANG, BEAU COKER (2020), "The age of secrecy and unfairness in recidivism prediction". In *Harvard Data Science Review*, 2 (1). https://doi.org/10.1162/99608f92.6ed64b30.
RYAN, N.J. (2018), "Five kinds of cyber deterrence". In *Philosophy & Technology*, 31, pp. 331-338. https://doi.org/10.1007/s13347-016-0251-1.
SALGANIK, MATTHEW J., IAN LUNDBERG, ALEXANDER T. KINDEL, CAITLIN E. AHEARN, KHALED AL-GHONEIM, ABDULLAH ALMAATOUQ, DREW M. ALTSCHUL, JENNIE E. BRAND, NICOLE BOHME CARNEGIE, RYAN JAMES COMPTON (2020), "Measuring the predictability of life outcomes with a scientific mass collaboration". In *Proceedings of the National Academy of Sciences*, 117 (15), pp. 8398-8403.
SAMUEL, ARTHUR L. (1960), "Some moral and technical consequences of automation – A refutation". In *Science*, 132 (3429), pp. 741-742. https://doi.org/10.1126/science.132.3429.741.
SARANTITIS, GEORGE (2020), "Data shift in machine learning: What is it and how to detect it". In *Georgios Sarantitis* (blog), 16 aprile. https://gsarantitis.wordpress.com/2020/04/16/data-shift-in-machine-learning-what-is-it-and-how-to-detect-it/.
SARTORIO, CAROLINA (2007), "Causation and responsibility". In *Philosophy Compass*, 2 (5), pp. 749-765. https://doi.org/10.1111/j.1747-9991.2007.00097.x.
SAVAS, ONUR, LEI DING, TERESA PAPALEO, IAN MCCULLOH (2020), "Adversarial attacks and countermeasures against ML models in army multi-domain operations". In *Artificial Intelligence and Machine Learning for Multi-domain Operations Applications II*, 11413, pp. 235-240. SPIE. https://doi.org/10.1117/12.2548798.
SCHELLING, THOMAS C. (1966), *Arms and Influence*. Yale University Press, New Haven, CT (*La diplomazia della violenza*. Tr. it. il Mulino, Bologna 1968).
SCHELLING, THOMAS C. (1980), *The Strategy of Conflict*. Harvard University Press, Cambridge, MA (*La strategia del conflitto*. Tr. it. Bruno Mondadori, Milano 2008).
SCHERRER, NINO, OLEXA BILANIUK, YASHAS ANNADANI, ANIRUDH GOYAL, PATRICK SCHWAB, BERNHARD SCHÖLKOPF, MICHAEL C. MOZER, YOSHUA BENGIO, STEFAN BAUER, NAN ROSEMARY KE (2022), "Learning neural causal models with active interventions". In *arXiv:2109.02429 [Cs, Stat]*, marzo. http://arxiv.org/abs/2109.02429.
SCHMITT, MICHAEL N. (2013), "Cyberspace and international law: The penumbral mist of uncertainty". In *Harvard Law Review Forum*, 126 (176), pp. 176-180.
SCHMITT, MICHAEL N. (2017) (a cura di), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations: Prepared by the International Groups of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence*. 2ª ed. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781316822524.
SCHMITT, MICHAEL N., JEFFREY S. THURNHER (2012), "Out of the loop: Autonomous weapon systems and the law of armed conflict". In *Harvard National Security Journal*, 4 (2), pp. 231-281.
SCHNEIER, BRUCE (2017), "Why the NSA makes us more vulnerable to cyberattacks". In *Foreign Affairs*, 30 maggio. https://www.foreignaffairs.com/articles/2017-05-30/why-nsa-makes-us-more-vulnerable-cyberattacks.

SCHUBERT, JOHAN, JOEL BRYNIELSSON, MATTIAS NILSSON, PETER SVENMARCK (2018), "Artificial intelligence for decision support in command and control systems". 23rd International Command and Control Research & Technology Symposium "Multi-Domain C2".

SCHULZKE, MARCUS (2013), "Autonomous weapons and distributed responsibility". In *Philosophy & Technology*, 26 (2), pp. 203-219. https://doi.org/10.1007/s13347-012-0089-0.

SCHULZKE, MARCUS (2016), "The morality of remote warfare: Against the asymmetry objection to remote weaponry". In *Political Studies*, 64 (1), pp. 90-105. https://doi.org/10.1111/1467-9248.12155.

SCHWARTZ, PETER J., DANIEL V. O'NEILL, MEGHAN E. BENTZ, ADAM BROWN, BRIAN S. DOYLE, OLIVIA C. LIEPA, ROBERT LAWRENCE, RICHARD D. HULL (2020), "AI-enabled wargaming in the military decision making process". In *Artificial Intelligence and Machine Learning for Multi-domain Operations Applications II*, 11413, pp. 118-134. SPIE. https://doi.org/10.1117/12.2560494.

SCULLEY, D., GARY HOLT, DANIEL GOLOVIN, EUGENE DAVYDOV, TODD PHILLIPS, DIETMAR EBNER, VINAY CHAUDHARY, MICHAEL YOUNG, JEAN-FRANÇOIS CRESPO, DAN DENNISON (2015), "Hidden technical debt in machine learning systems". In *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html.

SECHSER, TODD S., NEIL NARANG, CAITLIN TALMADGE (2019), "Emerging technologies and strategic stability in peacetime, crisis, and war". In *Journal of Strategic Studies*, 42 (6), pp. 727-735. https://doi.org/10.1080/01402390.2019.1626725.

SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE (2018), "AI in the UK: Ready, willing and able?". House of Lords, London.

SEPPÄLÄ, AKSELI, TEEMU BIRKSTEDT, MATTI MÄNTYMÄKI (2021), "From ethical AI principles to governed AI". In *2021 ICIS Proceedings*, pp. 1-17. Association for Information Systems, Austin, TX.

SHARIF, MAHMOOD, SRUTI BHAGAVATULA, LUJO BAUER, MICHAEL K. REITER (2016), "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition". In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security – CCS'16*, pp. 1528-1540. Association for Computing Machinery, Wien. https://doi.org/10.1145/2976749.2978392.

SHARKEY, AMANDA (2019), "Autonomous weapons systems, killer robots and human dignity". In *Ethics and Information Technology*, 21 (2), pp. 75-87. https://doi.org/10.1007/s10676-018-9494-0.

SHARKEY, NOEL E. (2008), "Cassandra or false prophet of doom: AI robots and war". In *IEEE Intelligent Systems*, 23 (4), pp. 14-17.

SHARKEY, NOEL E. (2010), "Saying 'no!' to lethal autonomous targeting". In *Journal of Military Ethics*, 9 (4), pp. 369-383. https://doi.org/10.1080/15027570.2010.537903.

SHARKEY, NOEL E. (2012a), "Killing made easy: From joysticks to politics". In PATRICK LIN, KEITH ABNEY, GEORGE BEKEY (a cura di), *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, Cambridge, MA, pp. 111-128.

SHARKEY, NOEL E. (2012b), "The evitability of autonomous robot warfare". In *International Review of the Red Cross*, 94 (886), pp. 787-799. https://doi.org/10.1017/S1816383112000732.

SHARKEY, NOEL E. (2016), "Staying in the loop: Human supervisory control of weapons". In CLAUS KREβ, HIN-YAN LIU, NEHAL BHUTA, ROBIN GEIβ, SUSANNE BECK (a cura di), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, Cambridge, pp. 23-38. https://doi.org/10.1017/CBO9781316597873.002.

SHAW, TYLER, ADAM EMFIELD, ANDRE GARCIA, EWART DE VISSER, CHRIS MILLER, RAJA PARASURAMAN, LISA FERN (2010), “Evaluating the benefits and potential costs of automation delegation for supervisory control of multiple UAVs”. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54 (19), pp. 1498-1502. https://doi.org/10.1177/154193121005401930.

SHIH, ANDY, ARJUN SAWHNEY, JOVANA KONDIC, STEFANO ERMON, DORSA SADIGH (2021), “On the critical role of conventions in adaptive human-AI collaboration”. In *arXiv:2104.02871 [Cs]*, aprile. http://arxiv.org/abs/2104.02871.

SHOEMAKER, DAVID (2017) (a cura di), *Oxford Studies in Agency and Responsibility*, vol. 4. Oxford University Press, New York.

SHUE, HENRY (2008), “Concept wars”. In *Survival*, 50 (2), pp. 185-192.

SIMON-MICHEL, JEAN HUGUES (2014), “Report of the 2014 informal meeting of experts on Lethal Autonomous Weapons Systems (LAWS)”. CCW/MSP/2014/3. In *High Contracting Parties to the Geneva Convention at the United Nations*, vol. 16, n. 2014, pp. 1-5. https://undocs.org/pdf?symbol=en/ccw/msp/2014/3. Ultimo accesso ottobre 2022.

SINHA, AMAN, HONGSEOK NAMKOONG, JOHN DUCHI (2017), “Certifying some distributional robustness with principled adversarial training”. In *arXiv:1710.10571 [Cs, Stat]*, ottobre. http://arxiv.org/abs/1710.10571.

SKERKER, MICHAEL, DUNCAN PURVES, RYAN JENKINS (2020), “Autonomous weapons systems and the moral equality of combatants”. In *Ethics and Information Technology*, 22 (3), pp. 197-209. https://doi.org/10.1007/s10676-020-09528-0.

SPARROW, ROBERT (2007), “Killer robots”. In *Journal of Applied Philosophy*, 24 (1), pp. 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

SPARROW, ROBERT (2016), “Robots and respect: Assessing the case against autonomous weapon systems”. In *Ethics & International Affairs*, 30 (1), pp. 93-116. https://doi.org/10.1017/S0892679415000647.

STEINHOFF, UWE (2013), “Killing them safely: Extreme asymmetry and its discontents”. In BRADLEY JAY STRAWSER (a cura di), *Killing by Remote Control: The Ethics of an Unmanned Military*. Oxford University Press, Oxford, pp. 179-208. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199926121.001.0001/acprof-978019 9926121-chapter-9.

STERNER, ERIC (2011), “Retaliatory deterrence in cyberspace”. In *Strategic Studies Quarterly*, 5 (1), pp. 62-80.

STEVENS, TIM (2012), “A cyberwar of ideas? Deterrence and norms in cyberspace”. In *Contemporary Security Policy*, 33 (1), pp. 148-170. https://doi.org/10.1080/13523260.2012.659597.

STEVENS, TIM (2020), “Knowledge in the grey zone: AI and cybersecurity”. In *Digital War*, 1 (1), pp. 164-170. https://doi.org/10.1057/s42984-020-00007-w.

STEVENSON, RYAN A., JOSEPH A. MIKELS, THOMAS W. JAMES (2007), “Characterization of the affective norms for English words by discrete emotional categories”. In *Behavior Research Methods*, 39 (4), pp. 1020-1024. https://doi.org/10.3758/BF03192999.

STILGOE, JACK, RICHARD OWEN, PHIL MACNAGHTEN (2013), “Developing a framework for responsible innovation”. In *Research Policy*, 42 (9), pp. 1568-1580. https://doi.org/10.1016/j.respol.2013.05.008.

STOLTZ, CHRISTOPHER (2018), “Augmenting the AOR: EOD airman provides critical skillset to army forensics team”. Air Force. https://www.af.mil/News/Article-Display/Article/1581694/augmenting-the-aor-eod-airman-provides-critical-skillset-to-army-forensics-team/.

STOWERS, KIMBERLY, LISA L. BRADY, CHRISTOPHER MACLELLAN, RYAN WOHLEBER, EDUARDO SALAS (2021), “Improving teamwork competencies in human-machine teams:

Perspectives from team science". In *Frontiers in Psychology*, 12 (maggio). https://doi.org/10.3389/fpsyg.2021.590290.
STRAWSER, BRADLEY J. (2013), "Introduction: The moral landscape of unmanned weapons". In BRADLEY JAY STRAWSER (a cura di), *Killing by Remote Control: The Ethics of an Unmanned Military*. Oxford University Press, Oxford, pp. 3-24. https://doi.org/10.1093/acprof:oso/9780199926121.003.0001.
STRAWSON, PETER (1962), "Freedom and resentment". In *Proceedings of the British Academy*, 48, pp. 1-25. Oxford University Press, Oxford.
STUMBORG, MICHAEL, BECKY ROH (2021), "Dimensions of autonomous decision-making". CNA. https://www.cna.org/CNA_files/PDF/Dimensions-of-Autonomous-Decision-making.pdf?utmsource=Center+for+Security+and+Emerging+Technology&utm_campaign=1280c55e66-EMAIL_CAMPAIGN_2022_01_27_0211&utmmedium=email&utm_term=0_fcbacf8c3e-1280c55e66-438318142.
SVIZZERA (2016), "Informal working paper submitted by Switzerland: Towards a 'compliance-based' approach to LAWS". 30 marzo. In *Informal Meeting of Experts on Lethal Autonomous Weapons Systems*. Genève. https://www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2016/meeting-experts-laws/documents/Switzerland-compliance.pdf.
SWEENEY, LATANYA (2013), "Discrimination in online ad delivery". Social Science Research Network, Rochester, NY.
SZEGEDY, CHRISTIAN, WOJCIECH ZAREMBA, ILYA SUTSKEVER, JOAN BRUNA, DUMITRU ERHAN, IAN GOODFELLOW, ROB FERGUS (2013), "Intriguing properties of neural networks". In *arXiv:1312.6199 [Cs]*, dicembre. http://arxiv.org/abs/1312.6199.
TADDEO, MARIAROSARIA (2012a), "An analysis for a just cyber warfare". In CHRISTIAN CZOSSECK, RAIN OTTIS, KATHARINA ZIOLKOWSKI (a cura di), *Fourth International Conference of Cyber Conflict*. NATO CCD COE and IEEE Publication, pp. 209-218.
TADDEO, MARIAROSARIA (2012b), "Information warfare: A philosophical perspective". In *Philosophy and Technology*, 25 (1), pp. 105-120.
TADDEO, MARIAROSARIA (2013), "Cyber security and individual rights, striking the right balance". In *Philosophy & Technology*, 26 (4), pp. 353-356. https://doi.org/10.1007/s13347-013-0140-9.
TADDEO, MARIAROSARIA (2014a), "Just information warfare". In *Topoi*, aprile, pp. 213-224. https://doi.org/10.1007/s11245-014-9245-8.
TADDEO, MARIAROSARIA (2014b), "The struggle between liberties and authorities in the information age". In *Science and Engineering Ethics*, settembre, pp. 1125-1138. https://doi.org/10.1007/s11948-014-9586-0.
TADDEO, MARIAROSARIA (2016a), "On the risks of relying on analogies to understand cyber conflicts". In *Minds and Machines*, 26 (4), pp. 317-321. https://doi.org/10.1007/s11023-016-9408-z.
TADDEO, MARIAROSARIA (2016b), "The moral value of information and information ethics". In LUCIANO FLORIDI (a cura di), *The Routledge Handbook of Philosophy of Information*. Routledge, New York, pp. 90-105.
TADDEO, MARIAROSARIA (2017a), "Cyber conflicts and political power in information societies". In *Minds and Machines*, 27 (2), pp. 265-268. https://doi.org/10.1007/s11023-017-9436-3.
TADDEO, MARIAROSARIA (2017b), "Deterrence by norms to stop interstate cyber attacks". In *Minds and Machines*, 27 (3), pp. 387-392. https://doi.org/10.1007/s11023-017-9446-1.
TADDEO, MARIAROSARIA (2017c), "Trusting digital technologies correctly". In *Minds and Machines*, 27 (4), pp. 565-568. https://doi.org/10.1007/s11023-017-9450-5.

TADDEO, MARIAROSARIA (2018a), “How to deter in cyberspace”. In *European Centre of Excellence for Countering Hybrid Threats*, 2018 (6), pp. 1-10.
TADDEO, MARIAROSARIA (2018b), “Deterrence and norms to foster stability in cyberspace”. In *Philosophy & Technology*, 31 (3), pp. 323-329. https://doi.org/10.1007/s13347-018-0328-0.
TADDEO, MARIAROSARIA (2018c), “The limits of deterrence theory in cyberspace”. In *Philosophy & Technology*, 31 (3), pp. 339-355. https://doi.org/10.1007/s13347-017-0290-2.
TADDEO, MARIAROSARIA (2020), “The ethical governance of the digital during and after the COVID-19 pandemic”. In *Minds and Machines*, 30 (2), pp. 171-176. https://doi.org/10.1007/s11023-020-09528-5.
TADDEO, MARIAROSARIA (2022), “A comparative analysis of the definitions of autonomous weapons systems”. In *Science and Engineering Ethics*, 28 (5), p. 37. https://doi.org/10.1007/s11948-022-00392-3.
TADDEO, MARIAROSARIA, ALEXANDER BLANCHARD (2022), “Accepting moral responsibility for the actions of autonomous weapons systems – A moral gambit”. In *Philosophy & Technology*, 35 (3), p. 78. https://doi.org/10.1007/s13347-022-00571-x.
TADDEO, MARIAROSARIA, ALEXANDER BLANCHARD, CHRIS THOMAS (2023), “From AI ethics principles to practices: A teleological methodology to apply AI ethics principles in the defence domain”. In *SSRN Electronic Journal*, 25 giugno. https://doi.org/10.2139/ssrn.4520945.
TADDEO, MARIAROSARIA, LUCIANO FLORIDI (2015), “The debate on the moral responsibilities of online service providers”. In *Science and Engineering Ethics*, novembre. https://doi.org/10.1007/s11948-015-9734-1.
TADDEO, MARIAROSARIA, LUCIANO FLORIDI (2018a), “Regulate artificial intelligence to avert cyber arms race”. In *Nature*, 556 (7701), pp. 296-298. https://doi.org/10.1038/d41586-018-04602-6.
TADDEO, MARIAROSARIA, LUCIANO FLORIDI (2018b), “How AI can be a force for good”. In *Science*, 361 (6404), pp. 751-752. https://doi.org/10.1126/science.aat5991.
TADDEO, MARIAROSARIA, LUDOVICA GLORIOSO (2016a) (a cura di), *Ethics and Policies for Cyber Operations*. Springer, New York.
TADDEO, MARIAROSARIA, LUDOVICA GLORIOSO (2016b), “Regulating cyber conflicts and shaping information societies”. In MARIAROSARIA TADDEO, LUDOVICA GLORIOSO (a cura di), *Ethics and Policies for Cyber Operations*. Springer, Berlin, pp. I-XVII.
TADDEO, MARIAROSARIA, TOM MCCUTCHEON, LUCIANO FLORIDI (2019), “Trusting artificial intelligence in cybersecurity is a double-edged sword”. In *Nature Machine Intelligence*, 1 (12), pp. 557-560. https://doi.org/10.1038/s42256-019-0109-1.
TADDEO, MARIAROSARIA, DAVID MCNEISH, ALEXANDER BLANCHARD, ELIZABETH EDGAR (2021), “Ethical principles for artificial intelligence in national defence”. In *Philosophy & Technology*, 34 (4), pp. 1707-1729. https://doi.org/10.1007/s13347-021-00482-3.
TADDEO, MARIAROSARIA, MARTA ZIOSI, ANDREAS TSAMADOS, LUCA GILLI, SHALINI KURAPATI (2022), “Artificial intelligence for national security: The predictability problem”. Centre for Emerging Technology and Security, London.
TALEB, NASSIM NICHOLAS (2007), *The Black Swan: The Impact of the Highly Improbable*. Random House, New York (*Il cigno nero. Come l’improbabile governa la nostra vita*. Tr. it. il Saggiatore, Milano 2008).
TAMBURRINI, GUGLIELMO (2016), “On banning autonomous weapons systems: From deontological to wide consequentialist reasons”. In BHUTA NEHAL, SUSANNE BECK, ROBIN GEIβ, HIN-YAN LIU, CAUS KREβ (a cura di), *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press, Cambridge, pp. 122-142.

TANJI, MICHAEL (2009), “Deterring a cyber attack? Dream on…”. In *Wired*, 19 febbraio. https://www.wired.com/2009/02/deterring-a-cyb/.

TAYLOR, ISAAC (2020), “Who is responsible for killer robots? Autonomous weapons, group agency, and the military-industrial complex”. In *Journal of Applied Philosophy*, 38, pp. 320-334. https://doi.org/10.1111/japp.12469.

TERZIS, PETROS (2020), “Onward for the freedom of others: Marching beyond the AI ethics”. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Barcelona, pp. 220-229. https://doi.org/10.1145/3351095.3373152.

“THE UK AND INTERNATIONAL HUMANITARIAN LAW 2018” (s.d.). https://www.gov.uk/government/publications/international-humanitarian-law-and-the-uk-government/uk-and-international-humanitarian-law-2018. Ultimo accesso 1º novembre 2020.

TIWARI, SAKSHI (2023), “Russia threatens to unleash ‘combat robot’ to burn Ukraine’s US & German-origin Abrams & Leopard 2 tanks”. In *Eurasian Times*, 7 gennaio. https://www.eurasiantimes.com/russia-threatens-to-unleash-combat-robot-to-burn-ukraines-us/.

TOBIN, DONAL (2022), “What is data cleansing and why does it matter?”. In *Integrate. Io* (blog), 21 gennaio. https://www.integrate.io/blog/what-does-data-cleansing-entail-and-why-does-it-matter/.

TOSSELL, CHAD, BOYOUNG KIM, BIANCA DONADIO, EWART DE VISSER, RYAN HOLEC, ELIZABETH PHILLIPS (2020), “Appropriately representing military tasks for human-machine teaming research”. In *Lecture Notes in Computer Science*, 12428, pp. 245-265. https://doi.org/10.1007/978-3-030-59990-4_19.

TSAMADOS, ANDREAS, NIKITA AGGARWAL, JOSH COWLS, JESSICA MORLEY, HUW ROBERTS, MARIAROSARIA TADDEO, LUCIANO FLORIDI (2021), “The ethics of algorithms: Key problems and solutions”. In *AI & Society*, febbraio. https://doi.org/10.1007/s00146-021-01154-8.

TSAMADOS, ANDREAS, LUCIANO FLORIDI, MARIAROSARIA TADDEO (2023), “The cybersecurity crisis of artificial intelligence: Unrestrained adoption and natural language-based attacks”. In *SSRN Electronic Journal*, 20 settembre. https://doi.org/10.2139/ssrn.4578165.

TSAMADOS, ANDREAS, MARIAROSARIA TADDEO (2023), “Human control of artificial intelligent systems: A critical review of key challenges and approaches”. In *SSRN Electronic Journal*, 9 luglio. https://doi.org/10.2139/ssrn.4504855.

UESATO, JONATHAN, BRENDAN O’DONOGHUE, AARON VAN DEN OORD, PUSHMEET KOHLI (2018), “Adversarial risk and the dangers of evaluating against weak attacks”. In *arXiv:1802.05666 [Cs, Stat]*, febbraio. http://arxiv.org/abs/1802.05666.

UK GOVERNMENT (2014), “Deterrence in the twenty-first century: Government response to the Committee’s Eleventh Report”. http://www.publications.parliament.uk/pa/cm201415/cmselect/cmdfence/525/52504.htm.

UK GOVERNMENT (2015), “National Security Strategy 2016-2021”. HM Government, London. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf.

UN GGE CCW (2019), *Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System (2019)*. Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. The United Nations Office at Geneva, Genève.

UN Institute for Disarmament Research (2014), “Cyber stability seminar 2014: Preventing cyber conflict”. https://unidir.org/wp-content/uploads/2023/05/cyber-stability-seminar-2014-en-612.pdf. Ultimo accesso luglio 2024.

UNIDIR (United Nations Institute for Disarmament Research) (2017), "The weaponization of increasingly autonomous technologies: Concerns, characteristics and definitional approaches". UNIDIR Resources.

UNITED NATIONS HIGH COMMISSIONER FOR HUMAN RIGHTS (2014), *The Right to Privacy in the Digital Age: Annual Report of the United Nations High Commissioner for Human Rights and Reports of the Office of the High Commissioner and the Secretary-General*. A/HRC/27/37. United Nations Human Rights Council, Genève.

UNITED NATIONS HIGH COMMISSIONER FOR HUMAN RIGHTS (2021), T*he Right to Privacy in the Digital Age: Annual Report of the United Nations High Commissioner for Human Rights and Reports of the Office of the High Commissioner and the Secretary-General*. A/HRC/48/31. United Nations Human Rights Council, Genève.

US ARMY (2017), "Robotic and autonomous systems strategy". https://www.tradoc.army.mil/Portals/14/Documents/RAS_Strategy.pdf.

US DEPARTMENT OF DEFENSE (2012), "DoD Directive 3000.09 on autonomy in weapon systems". https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

US DEPARTMENT OF DEFENSE (2022a), "Fact sheet on U.S. security assistance for Ukraine". 10 maggio. https://www.defense.gov/News/Releases/Release/Article/3027295/fact-sheet-on-us-security-assistance-for-ukraine/.

US DEPARTMENT OF DEFENSE (2022b), "Responsible artificial intelligence strategy and implementation pathway".

US GOVERNMENT (2015), "The Department of Defense Cyber Strategy". Washington, DC.

US SENATE SELECT COMMITTEE ON INTELLIGENCE (2002), *Joint Inquiry into Intelligence Community Activities before and after the Terrorist Attacks of September 11, 2001*. U.S. Senate Select Committee on Intelligence and U.S. House Permanent Select Committee on Intelligence, Washington, DC.

VEERAMACHANENI, KALYAN, IGNACIO ARNALDO, ALFREDO CUESTA-INFANTE, VAMSI KORRAPATI, COSTAS BASSIAS, KE LI (2016), "AI^2: Training a big data machine to defend". In 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), p. 13. https://ieeexplore.ieee.org/document/7502263.

VERDIESEN, ILSE, FILIPPO SANTONI DE SIO, VIRGINIA DIGNUM (2021), "Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight". In *Minds and Machines*, 31 (1), pp. 137-163. https://doi.org/10.1007/s11023-020-09532-9.

VOGEL, KATHLEEN M., GWENDOLYNNE REID, CHRISTOPHER KAMPE, PAUL JONES (2021), "The impact of AI on intelligence analysis: Tackling issues of collaboration, algorithmic transparency, accountability, and management". In *Intelligence and National Security*, 36 (6), pp. 827-848.

WAGNER, MARKUS (2014), "The dehumanization of international humanitarian law: Legal, ethical, and political implications of autonomous weapon systems". In *Vanderbilt Journal of Transnational Law*, 47, pp. 1371-1424.

WALCH, KATHLEEN (2020), "How AI is finding patterns and anomalies in your data". In *Forbes*, 10 maggio. https://www.forbes.com/sites/cognitiveworld/2020/05/10/finding-patterns-and-anomalies-in-your-data/.

WALLACE, R. JAY (1998), *Responsibility and the Moral Sentiments*. Harvard University Press, Cambridge, MA.

WALLISER, JAMES C., EWART J. DE VISSER, EVA WIESE, TYLER H. SHAW (2019), "Team structure and team building improve human-machine teaming with autonomous agents". In

Journal of Cognitive Engineering and Decision Making, 13 (4), pp. 258-278. https://doi.org/10.1177/1555343419867563.

WALZER, MICHAEL (1977), *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books, New York (*Guerre giuste e ingiuste. Un discorso morale con esemplificazioni storiche*. Tr. it. Laterza, Roma-Bari 2009).

WALZER, MICHAEL (2006), *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. 4ª ed. Basic Books, New York.

WATSON, GARY (1975), "Free agency". In *Journal of Philosophy*, 72 (8), pp. 205-220. https://doi.org/10.2307/2024703.

WEERAMANTRY, C.G. (1985), "Nuclear weaponry and scientific responsibility". In *Journal of the Indian Law Institute*, 27 (3), pp. 351-386.

WEINBAUM, CORTNEY, JOHN N.T. SHANAHAN (2018), "Intelligence in a data-driven age". In *Joint Force Quarterly*, 90, pp. 4-9.

WHITTLESTONE, JESS, RUNE NYRUP, ANNA ALEXANDROVA, STEPHEN CAVE (2019), "The role and limits of principles in AI ethics: Towards a focus on tensions". In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, Honolulu, HI, pp. 195-200. https://doi.org/10.1145/3306618.3314289.

WIDDERSHOVEN, GUY, TINEKE ABMA, BERT MOLEWIJK (2009), "Empirical ethics as dialogical practice". In *Bioethics*, 23 (4), pp. 236-248. https://doi.org/10.1111/j.1467-8519.2009.01712.x.

WIENER, NORBERT (1960), "Some moral and technical consequences of automation". In *Science*, 131 (3410), pp. 1355-1358. https://doi.org/10.1126/science.131.3410.1355.

WINTER, ELLIOT (2018), "Autonomous weapons in humanitarian law: Understanding the technology, its compliance with the principle of proportionality and the role of utilitarianism". In *Groningen Journal of International Law*, 6 (1), pp. 183-202.

WINTER, ELLIOT (2020), "The compatibility of autonomous weapons with the principles of distinction in the law of armed conflict". In *International & Comparative Law Quarterly*, 69 (4), pp. 845-876.

WITTGENSTEIN, LUDWIG (2009), *Philosophical Investigations*. 4ª ed. Tr. ing. di G.E.M. Anscombe. Wiley-Blackwell, Malden, MA (*Ricerche filosofiche*. Tr. it. Feltrinelli, Milano 2024).

WOODBURY, MARSHA (2003), *Computer and Information Ethics*. Stipes, Champaign, IL.

WOODS, DAVID D., EMILY S. PATTERSON, EMILIE M. ROTH (2002), "Can we ever escape from data overload? A cognitive systems diagnosis". In *Cognition, Technology & Work*, 4 (1), pp. 22-36. https://doi.org/10.1007/s101110200002.

WOOLDRIDGE, MICHAEL J. (2020), *The Road to Conscious Machines: The Story of AI*. Pelican, London.

WOOLDRIDGE, MICHAEL J., NICHOLAS R. JENNINGS (1995), "Intelligent agents: Theory and practice". In *Knowledge Engineering Review*, 10 (2), pp. 115-152. https://doi.org/10.1017/S0269888900008122.

YANG, GUANG-ZHONG, JIM BELLINGHAM, PIERRE E. DUPONT, PEER FISCHER, LUCIANO FLORIDI, ROBERT FULL, NEIL JACOBSTEIN, ET AL. (2018), "The grand challenges of science robotics". In *Science Robotics*, 3 (14), eaar7650. https://doi.org/10.1126/scirobotics.aar7650.

YARON, MAYA (2018), "Statement by Maya Yaron to the Convention on Certain Conventional Weapons (CCW) GGE on Lethal Autonomous Weapons Systems (LAWS)". Permanent Mission of Israel to the UN, Genève. https://www.unog.ch/80256EDD006B8954/(httpAssets)/990162020E17A5C9C12582720057E720/$file/2018_LAWS6bIsrael.pdf.

YOU, SANGSEOK, LIONEL ROBERT (2016), “Emotional attachment, performance, and viability in teams collaborating with Embodied Physical Action (EPA) robots”. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1810&context=jais.

ZAGARE, FRANK C., D. MARC KILGOUR (2000), *Perfect Deterrence*. Cambridge University Press, Cambridge.

ZEGART, AMY B. (2005), “September 11 and the adaptation failure of U.S. intelligence agencies”. In *International Security*, 29 (4), pp. 78-111.

ZEGART, AMY B. (2022), *Spies, Lies, and Algorithms: The History and Future of American Intelligence*. Princeton University Press, Princeton, NJ.

ZHUGE, JIANWEI, THORSTEN HOLZ, XINHUI HAN, CHENGYU SONG, WEI ZOU (2007), “Collecting autonomous spreading malware using high-interaction honeypots”. In SIHAN QING, HIDEKI IMAI, GUILIN WANG (a cura di), *Information and Communications Security*. Springer, Berlin-Heidelberg, pp. 438-451.

SCIENZA E IDEE

Ultimi volumi pubblicati

180. P. Sloterdijk, *Il furore di Dio*
181. J. Gribbin, *L'Universo. Una biografia*
182. A. Lancelin, M. Lemonnier, *I filosofi e l'amore*
183. E. Bellone, *Molte nature*
184. M. Detienne, *I giardini di Adone*
185. C. Frith, *Inventare la mente*
186. G. Gigerenzer, *Decisioni intuitive*
187. R. Brague, *Il Dio dei cristiani*
188. P. Hadot, *Ricordati di vivere*
189. G. Brown, *Una vita senza fine?*
190. I. Tattersall, *Il mondo prima della storia*
191. M.S. Gazzaniga, *Human*
192. S. Dehaene, *I neuroni della lettura*
193. M. Gribbin, J. Gribbin, *Cacciatori di piante*
194. P. Goodchild, *Il vero dottor Stranamore*
195. M. Tomasello, *Le origini della comunicazione umana*
196. K. Jaspers, *Introduzione alla filosofia*
197. T. Metzinger, *Il tunnel dell'io*
198. G. Guidorizzi, *Ai confini dell'anima*
199. M. Romano, *Ascesa e declino della città europea*
200. E. Bencivenga, *La filosofia come strumento di liberazione*
201. A.D. Aczel, *Le cattedrali della preistoria*
202. A. Noë, *Perché non siamo il nostro cervello*
203. R.L. Gregory, *Vedere attraverso le illusioni*
204. J.R. Searle, *Creare il mondo sociale*
205. P. Sloterdijk, *Devi cambiare la tua vita*

206. R. Scruton, *Bevo dunque sono*
207. S.S. Gubser, *Il piccolo libro delle stringhe*
208. C. de Duve, *Genetica del peccato originale*
209. S. Dehaene, *Il pallino della matematica*
210. W.G. Naphy, *La rivoluzione protestante*
211. G. Boniolo, *Il pulpito e la piazza*
212. N. Carr, *Internet ci rende stupidi?*
213. C. de Seta, *Il fascino dell'Italia nell'età moderna*
214. V. Bianchi, *Dracula*
215. T. Pievani, *La vita inaspettata*
216. S. Tagliagambe, A. Malinconico, *Pauli e Jung*
217. K.A. Appiah, *Il codice d'onore*
218. R. Dunbar, *Di quanti amici abbiamo bisogno?*
219. T. Nathan, *Una nuova interpretazione dei sogni*
220. S. Kirshenbaum, *La scienza del bacio*
221. P. Hadot, *La felicità degli antichi*
222. P. Brusasco, *Babilonia*
223. C. Bartocci, *Una piramide di problemi*
224. F. Close, *Neutrino*
225. P.S. Churchland, *Neurobiologia della morale*
226. S. Baron-Cohen, *La scienza del male*
227. J.K. O'Regan, *Perché i colori non suonano*
228. F.R. Young, *Bolle, gocce, schiume*
229. A. Desmond, J. Moore, *La sacra causa di Darwin*
230. E.O. Wilson, *La conquista sociale della Terra*
231. G. Farmelo, *L'uomo più strano del mondo*
232. G. Guidorizzi, *Il compagno dell'anima*
233. H. Rheingold, *Perché la rete ci rende intelligenti*
234. A. Portmann, *La forma degli animali*
235. F. Conti, *Claude Bernard e la nascita della biomedicina*
236. F. de Waal, *Il bonobo e l'ateo*
237. D. Chamovitz, *Quel che una pianta sa*

238. R. Dunbar, *Amore e tradimento*
239. R. Casati, *Dov'è il Sole di notte?*
240. J.-P. Changeux, *Il bello, il buono, il vero*
241. L.A. Sass, *Follia e modernità*
242. E.O. Wilson, *Lettere a un giovane scienziato*
243. S.M. Aglioti, G. Berlucchi, *Neurofobia*
244. C. Rovelli, *La realtà non è come ci appare*
245. D. Maestripieri, *A che gioco giochiamo noi primati*
246. D.C. Dennett, *Strumenti per pensare*
247. P.S. Churchland, *L'io come cervello*
248. D. Le Breton, *Esperienze del dolore*
249. S. Dehaene, *Coscienza e cervello*
250. D. Edmonds, *Uccideresti l'uomo grasso?*
251. A. Norenzayan, *Grandi Dei*
252. C. Bartocci, *Dimostrare l'impossibile*
253. A.D. Aczel, *Perché la scienza non nega Dio*
254. G. Cosmacini, *Medicina e rivoluzione*
255. G. Gigerenzer, *Imparare a rischiare*
256. S. Bocchi, *Zolle*
257. N. Carr, *La gabbia di vetro*
258. L. Susskind, A. Friedman, *Meccanica quantistica*
259. T. Nagel, *Mente e cosmo*
260. A. Gefter, *Due intrusi nel mondo di Einstein*
261. S. LeVay, *Gay si nasce?*
262. G. Madhavan, *Come pensano gli ingegneri*
263. V. Gallese, M. Guerra, *Lo schermo empatico*
264. B. Nassim Aboudrar, *Come il velo è diventato musulmano*
265. A.D. Aczel, *Caccia allo zero*
266. J. LeDoux, *Ansia*
267. P. Halpern, *I dadi di Einstein e il gatto di Schrödinger*
268. G. Lolli, *Tavoli, sedie, boccali di birra*
269. J.R. Searle, *Vedere le cose come sono*

270. F. de Waal, *Siamo così intelligenti da capire l'intelligenza degli animali?*
271. D. Le Breton, *Fuggire da sé*
272. M.C. Corballis, *La mente che vaga*
273. G. Giorello, E. Sindoni, *Un mondo di mondi*
274. M. Tomasello, *Storia naturale della morale umana*
275. A. Benini, *Neurobiologia del tempo*
276. C. Rugarli, *Medici a metà*
277. R. Menzel, M. Eckoldt, *L'intelligenza delle api*
278. M. Simon, *La vespa che fece il lavaggio del cervello al bruco*
279. L. Floridi, *La quarta rivoluzione*
280. M. Lewis, *Un'amicizia da Nobel*
281. A. Moro, *Le lingue impossibili*
282. M. Malvaldi, *L'architetto dell'invisibile*
283. G. Guidorizzi, *I colori dell'anima*
284. G. Segrè, B. Hoerlin, *Il Papa della fisica*
285. S. Sloman, P. Fernbach, *L'illusione della conoscenza*
286. N. deGrasse Tyson, *Astrofisica per chi va di fretta*
287. G.M. Edelman, *Darwinismo neurale*
288. M. Tegmark, *Vita 3.0*
289. D.C. Dennett, *Dai batteri a Bach*
290. A. Benini, *La mente fragile*
291. D. Le Breton, *Sul silenzio*
292. E. Boncinelli, *La farfalla e la crisalide*
293. H. Collins, *Un bacio tra le stelle*
294. P. Sloterdijk, *Dopo Dio*
295. E.O. Wilson, *Le origini della creatività*
296. L. Susskind, A. Friedman, *Relatività ristretta e teoria classica dei campi*
297. E.R. Kandel, *La mente alterata*
298. F. Noudelmann, *Il genio della menzogna*
299. P.A.M. Dirac, *La bellezza come metodo*

300. J.R. Searle, *Il mistero della realtà*
301. G. Lipovetsky, *Piacere e colpire*
302. G. Rizzolatti, C. Sinigaglia, *Specchi nel cervello*
303. M.S. Gazzaniga, *La coscienza è un istinto*
304. P. Caraveo, *Conquistati dalla Luna*
305. A. Touraine, *In difesa della modernità*
306. G. Kepel, *Uscire dal caos*
307. D. Reich, *Chi siamo e come siamo arrivati fin qui*
308. M. Dorato, *Disinformazione scientifica e democrazia*
309. M. Tomasello, *Diventare umani*
310. D. Le Breton, *Ridere*
311. S. Hossenfelder, *Sedotti dalla matematica*
312. S. Dehaene, *Imparare*
313. L. Susskind, G. Hrabovsky, *Il minimo teorico*
314. L. Floridi, *Pensare l'infosfera*
315. A. Piontelli, *Il culto del feto*
316. F. de Waal, *L'ultimo abbraccio*
317. J. Derrida, *Politiche dell'amicizia*
318. E.O. Wilson, *Le origini profonde delle società umane*
319. R. Panek, *Il mistero sotto i nostri piedi*
320. S. Rossi, *Il cervello elettrico*
321. A. Cairo, *Come i grafici mentono*
322. A. Aguirre, *Zen e multiversi*
323. A. Blum, *Rosso di sera…*
324. J. LeDoux, *Lunga storia di noi stessi*
325. R. Simone, *Il software del linguaggio*
326. M. Malvaldi, *La direzione del pensiero*
327. A.C.A. Elliott, *È grande questo numero?*
328. C. Koch, *Sentirsi vivi*
329. D. Dorling, *Rallentare*
330. D. Coen, *L'arte della probabilità*
331. S. Baron-Cohen, *I geni della creatività*

332. K. Davies, *Riscrivere l'umanità*
333. E.O. Wilson, *Storie dal mondo delle formiche*
334. M. Schilthuizen, *Darwin va in città*
335. F. Ciceri, P. Arosio, *Come batteremo il cancro*
336. G. Ravasi, *Biografia di Gesù*
337. F. Brevini, *Abbiamo ancora bisogno degli intellettuali?*
338. S. Weidensaul, *In volo sul mondo*
339. D.T. Blumstein, *Paura*
340. L. Floridi, *Etica dell'intelligenza artificiale*
341. D.C. Dennett, G.D. Caruso, *A ognuno quel che si merita*
342. S.M. Fleming, *Conoscere se stessi*
343. L.M. Krauss, *La fisica del cambiamento climatico*
344. A. Benini, *Neurobiologia della volontà*
345. F. de Waal, *Diversi*
346. O. Sibony, *Stai per commettere un terribile errore!*
347. S. Nadler, L. Shapiro, *Quando persone intelligenti hanno idee stupide*
348. P. Odifreddi, *Pillole matematiche*
349. A. Colombo, *Il governo mondiale dell'emergenza*
350. M. Tomasello, *Dalle lucertole all'uomo*
351. G. Gigerenzer, *Perché l'intelligenza umana batte ancora gli algoritmi*
352. A. Seth, *Come il cervello crea la nostra coscienza*
353. P. Ferri, *Le sfide di Marte*
354. M. Walzer, *Che cosa significa essere liberale*
355. F. Pregliasco, P. Arosio, *I superbatteri*
356. D.J. Chalmers, *Più realtà*
357. D.C. Dennett, *Coscienza*
358. P. Odifreddi, *A piccole dosi*
359. M. Galli, *Una banale influenza?*
360. N. deGrasse Tyson, D. Goldsmith, *Origini*
361. S. Rossi, D. Prattichizzo, *Il corpo artificiale*
362. P. Burke, *Ignoranza*
363. E. Selya, *Salvador Luria*

364. L. Susskind, A. Cabannes, *Relatività generale*
365. B. Larsson, *Essere o non essere umani*
366. J. LeDoux, *I quattro mondi dell'uomo*
367. S. Talamo, *Misurare la storia*
368. T. Sharot, C.R. Sunstein, *Guardate meglio*
369. L. Floridi, *Filosofia dell'informazione*
370. G. Ravasi, *Ero un blasfemo, un persecutore e un violento*
371. S. Carroll, *Spazio, tempo, movimento*
372. D.C. Dennett, *Pensandoci bene*
373. T.S. Kuhn, *L'incommensurabilità nella scienza*
374. P. Ferri, *Volare oltre il cielo*
375. G. Benenti, G. Casati, S. Montangero, *Il computer impossibile*
376. D. Quammen, *L'evoluzionista riluttante*
377. P. Odifreddi, *Incontri ravvicinati tra le due culture*
378. N. Carr, *Superbloom*
379. C.R. Sunstein, *Come diventare famosi*