CAMERA DEI DEPUTATI N. **4793**

PROPOSTA DI LEGGE

d'iniziativa del deputato QUINTARELLI

Delega al Governo per la disciplina e la promozione dello sviluppo dei sistemi di intelligenza artificiale

Presentata il 18 dicembre 2017

ONOREVOLI COLLEGHI ! — il Professor David Weinberger scrive « Dovremmo essere tutti preoccupati che i sistemi di intelligenza artificiale (AI) basati sul *machine learning*, cui affidiamo decisioni – da chi assumere a chi assicurare – possano commettere errori seri, nuocendo agli individui e ad intere categorie di persone. Alcune volte gli errori sono relativamente innocui, come ad esempio quando il riconoscitore di immagini di Google decide che mani e braccia umane sono elementi costitutivi di un bilanciere. È stato molto peggio, quando in USA *Google Photos* ha identificato uomini di colore come gorilla. È devastante quando, sempre in USA, una AI raccomanda che uomini di colore siano tenuti in carcere più a lungo degli uomini bianchi semplicemente in ragione della loro razza. E non si può non pensare ad errori gravissimi che possono essere realizzati da sistemi letali di arma autonomi (LAWS) che potrebbero ammantare il razzismo artificiale in mezzi volanti esplosivi ».

Anche per tentare di migliorare questa situazione di ingiustizie artificiali, il Parlamento europeo ha approvato il Regolamento generale sulla protezione dei dati (*General Data Protection Regulation* - GDPR), che è spesso usato per asserire un « diritto alla spiegazione » per algoritmi che incidono in modo significativo sugli utenti. Ma asserire in modo generale un simile diritto, ovvero che una intelligenza artificiale debba essere in grado di spiegare la motivazione con cui giunge a una determinazione, può essere una sfida tecnica rilevante o persino inattuabile. Affermare un simile diritto in senso generale può indurre un rallentamento nello sviluppo e nell'applicazione dell'intelligenza artificiale per benefici usi civili. Nessuna spiegazione ha un costo nullo.

Se, ad esempio, il nostro medico curante ci informasse che un sistema di diagnostica basato sul *deep learning* ha concluso che c'è

una forte probabilità che nei prossimi due anni svilupperemo una forma tumorale, vorremmo sapere le motivazioni di questa affermazione. Un tale sistema potrebbe essere stato addestrato sulle informazioni sanitarie di 70.000 pazienti e aver riscontrato in esse migliaia di fattori che possono contribuire a un calcolo probabilistico. Il medico non potrebbe mai rispondere in modo accurato a una richiesta di motivazioni: ci proporrebbe delle possibili opzioni terapeutiche basate su questa analisi e includere l'opzione di ignorare le conclusioni dell'intelligenza artificiale.

Sarebbe un comportamento assai diverso da ciò che normalmente ci aspetteremmo da un essere umano che fornisce una spiegazione. Normalmente il medico ci direbbe che un certo tipo di tumore è spesso determinato da talune cause e che sulla base di determinate analisi c'è un riscontro di tali cause; pertanto è probabile che il tumore si manifesti e conseguentemente il medico ci presenterebbe le opzioni per minimizzare il rischio.

Ma i sistemi di *machine learning* non devono attenersi ai modelli umani di esplicitazione di come i vari fattori si correlano. Indubbiamente in un sistema si introducono fattori che polarizzano le determinazioni, a causa della decisione dei *dataset* di addestramento, dei *bias* impliciti dei programmatori dei sistemi, dell'interfaccia utente con cui gli utenti operano sul sistema eccetera. Ma il sistema non è conformato su come riteniamo che i dati debbano relazionarsi. Il sistema stesso compie interazioni di analisi dei dati e trova correlazioni probabilistiche complesse che divengono parte del modello che il sistema si costruisce.

I modelli costruiti dagli umani tendono a ridurre le variabili a un insieme sufficientemente piccolo, identificando quelle più significative, in modo tale da renderle semplicemente analizzabili e comprensibili. Ma i modelli di *machine learning* non presentano le stesse limitazioni. Possono produrre modelli molto complessi che operano in modo efficace, ad esempio per diagnosticare quadri clinici, ma che non possono essere ridotti e semplificati a un livello che li renda comprensibili o « leggibili » da un essere umano.

Alcune volte l'assenza di spiegabilità non è rilevante. Ad esempio, i sistemi di raccomandazione di film o libri, non incidendo in modo significativo sulle persone, possono essere tranquillamente esentati dalla spiegazione delle motivazioni delle conclusioni.

Non abbiamo neppure bisogno di comprendere come un sistema di *machine learning* operi per poterne correggere il funzionamento. Se si rileva che un sistema di raccomandazione di prodotti non genera un aumento dei ricavi può essere semplicemente riaddestrato su dati diversi o calibrato misurando gli incrementi di efficacia. Se questa attività produce un aumento dei ricavi, una spiegazione può essere interessante ma non è necessaria.

Ma immaginiamo che un sistema di intelligenza artificiale che interviene nella selezione del personale ci respinga a causa della nostra età, genere, razza eccetera. Indubbiamente vorremmo avere conoscenza delle motivazioni, come prevede il GDPR. Ma, per quanto esula dalle situazioni previste dal GDPR, le spiegazioni non sono l'unico mezzo di verifica. In alcuni casi potremmo verificare se una decisione è stata presa basandosi su variabili non ammissibili, come ad esempio razza o sesso, senza avere una spiegazione completa delle motivazioni. Ciò potrebbe avvenire con un semplice *test* riproponendo le stesse informazioni di profilo, modificando gli *input* relativi alle sole variabili non ammissibili e valutando *l'output*.

Anche se eseguissimo tale *test* – potremmo non essere in grado di « riavvolgere » il sistema usando dati complessi in tempo reale – potremmo non fidarci dei risultati che ci indicano che non c'è stato un *bias* nelle decisioni in quanto il sistema potrebbe inavvertitamente usare dei collegamenti indiretti non evidenti, ad esempio per determinare il sesso, che noi non potremmo individuare. Le variabili potrebbero essere così complesse che potremmo non capire che esse consentono di determinare il sesso, così come non riusciamo a capire perché un insieme complesso di va-

riabili indica una predisposizione a sviluppare un determinato tipo di tumore. In ogni caso, scoprire che il sesso influisce nella determinazione dell'*output* non necessariamente indica che ciò abbia avuto un ruolo determinante nella decisione, in quanto altre migliaia di variabili potrebbero averlo determinato. Sarebbe come dire che una persona si sente male per la pressione atmosferica, mentre la causa è aver mangiato cibo alterato, su una barca molto movimentata, nel caldo soffocante, con una nuova configurazione batterica nello stomaco, mentre c'era bassa pressione. Sapere che il sesso è stato un fattore determinante nel processo di selezione per un lavoro potrebbe essere sufficiente per richiedere qualche tipo di risarcimento, ma non costituisce in sé quello che noi intendiamo generalmente come « spiegazione » e può andare bene, perché una verifica non richiede necessariamente di generare una spiegazione.

Una verifica non è l'unico modo per gestire una richiesta di qualcuno che pensi di essere stato discriminato nella selezione per un posto di lavoro a causa del suo sesso. Possiamo anche verificare l'*output* complessivo per vedere se oltre il 50 per cento (o quale che sia il livello *standard* di « equità » concordato) di chi ha fatto richieste che sono state accolte siano donne ben qualificate. Nel caso si rilevi che non è così, possiamo tarare il sistema senza necessariamente capire come sia stato introdotto il *bias*. Se troviamo una spiegazione – ad esempio i *dataset* di addestramento incorporavano *bias* esistenti – non dovremmo necessariamente aver bisogno di una spiegazione perché tale o tale altra donna sia stata esclusa.

Per mantenere l'equità di un sistema, le spiegazioni sono meno rilevanti di quanto possa apparire a prima vista.

Quindi, possiamo vincolare l'intelligenza artificiale a qualcosa che una mente umana possa capire o possiamo consentire che questi sistemi aumentino in complessità progressivamente, mano a mano che trattano una maggiore quantità di dati sempre più complessi o che interagiscono con altri sistemi in altri settori adiacenti, come ad esempio un sistema di previsioni economiche che incorpora *output* di altri sistemi di intelligenza artificiale inspiegabili su dati sanitari e ambientali globali.

Se decideremo di consentire ai sistemi di intelligenza artificiale di svilupparsi oltre le possibilità di comprensione umana avremo bisogno di altri strumenti per raggiungere gli obiettivi per cui abbiamo generalmente usato le spiegazioni.

Inquadrare queste discussioni in termini di ottimizzazione può consentirci di gestire aspetti etici con la grazia di un'auto autonoma che scarta per evitare un gruppo di suore che spinge dei passeggini.

Consideriamo un veicolo autonomo (VA) che percorra una strada cittadina chiusa a veicoli guidati da umani. Immaginiamo che accada un evento imprevedibile che non conceda al VA nessuna opzione positiva (un nugolo di cuccioli cade da un sovrappasso, una voragine si apre sotto un autobus pieno di geni matematici eccetera). Qualunque sia il dilemma, immaginiamo che l'opzione « meno peggiore » per la rete di VA sia di condurre il veicolo con zia Ada giù da un ponte. Affrontiamo questi esperimenti mentali per chiarire i princìpi morali che desideriamo siano seguiti dai VA. Questo è utile come approfondimento mentale, ma può condurrei in un binario morto nel momento in cui dobbiamo considerare come disegnarne e governiamo le intelligenze artificiali: esaminiamo gli algoritmi per vedere se hanno violato dei principi morali. Ciò ci ripropone il tema delle spiegazioni in quanto assumiamo che queste decisioni sono – o dovrebbero essere – governate dall'applicazione di princìpi. Spieghiamo le « decisioni » dei VA sulla base di princìpi applicati a determinate condizioni.

Ma nel mondo dell'intelligenza artificiale, così come nella vita quotidiana, ciò ci conduce ad argomentazioni sulla moralità che sono spesso impossibili da risolvere in quanto comprendiamo i princìpi riferendoli a casi generali che sono estremamente difficili da valutare nell'applicazione a casi particolari.

I princìpi dovrebbero essere incorporati *ex ante* in tutti i sistemi di intelligenza artificiale che produciamo, determinan-

done il loro comportamento senza eccezioni, mentre nei sistemi gestiti da umani le valutazioni sono fatte *ex post*, con decisioni valutate caso per caso nella loro specificità, senza così poter determinare regole generali con validità assolute.

La domanda relativa ai VA non può essere formulata *ex ante* in termini specifici, più di quanto avvenga per i comportamenti umani, se sia preferibile che scarti addosso a delle suore o all'intera famiglia Rossi. La domanda *ex ante* può riguardare come desideriamo che il sistema complessivo di intelligenza artificiale si comporti. Il sistema può comprendere principi deontologici ma fondamentalmente il VA prenderà « decisioni » secondo modalità diverse dall'applicazione di princìpi a casi particolari di pericolo. I sistemi di intelligenza artificiale prenderanno decisioni sulla base degli obiettivi per cui sono stati ottimizzati perché questo è il modo e la ragione per cui li costruiamo.

I sistemi di intelligenza artificiale dovrebbero dichiarare per cosa sono ottimizzati. Le ottimizzazioni di sistemi che incidono significativamente sui cittadini non dovrebbero essere decise dalle società che creano questi sistemi, ma da organizzazioni che rappresentino gli interessi pubblici.

L'ottimizzazione deve essere definita in termini di esiti rispetto a degli obiettivi.

Un sistema di VA è efficacemente ottimizzato per ridurre gli incidenti mortali se su un intervallo di tempo statisticamente significativo il numero degli incidenti mortali diminuisce. L'ottimizzazione non deve precisare un obiettivo specifico per misurare il successo. L'obiettivo è il massimo desiderabile e possibile, date tutte le ottimizzazioni desiderate e le vicissitudini del contesto.

I disegnatori di sistemi parlano di ottimizzazione perché riconoscono che le macchine sono imperfette e spesso disegnate per servire insiemi inconsistenti di obiettivi. Il tuo veicolo sarà ottimizzato per massima autonomia, impatto ambientale, prezzo, accelerazione, sicurezza, *comfort* o prestigio ? Preferire la sicurezza può richiedere un telaio più pesante con effetti negativi su autonomia, accelerazione e impatto ambientale. Privilegiare la riduzione di impatto ambientale può significare tempi di viaggio maggiori e minor *comfort*. I disegnatori devono creare un bilanciamento di ottimizzazioni, determinando quanto si sacrifica di un aspetto per migliorare una combinazione degli altri valori.

Come dice David P. Reed, uno degli architetti di *internet*, ottimizzare un sistema per un valore deottimizza gli altri.

Mentre le ottimizzazioni si applicano a sistemi, esse possono essere determinate, con alcuni limiti, dai singoli utenti. Ad esempio i passeggeri in un VA possono ottimizzare un viaggio per avere un migliore panorama. Muovere in alto questo ipotetico selettore – un giorno disponibile in un cruscotto – muoverà automaticamente verso il basso altri selettori: il viaggio potrebbe richiedere più tempo e più energia.

I limiti imposti alla possibilità degli utenti di modificare i selettori saranno determinati da coloro che disegnano il sistema di ottimizzazione nel suo complesso. In taluni casi gli utenti potranno non essere abilitati a ottimizzare un viaggio specifico in un modo che deottimizzi anche di poco il sistema complessivo per salvare vite.

Perché è rilevante discutere della *governance* dei sistemi di intelligenza artificiale in termini di ottimizzazione ?

In primo luogo, evade il dilemma costruzione *ex ante versus* valutazione *ex post* e ci focalizza su una discussione normativa laica sull'intelligenza artificiale come strumento per tendere a benefìci socialmente desiderabili (obiettivo della politica) invece di un enigma labirintico di discussioni su princìpi e loro applicazioni. Ad esempio, come società non riusciremo mai a convergere *ex ante* su una decisione se un VA deve essere programmato per investire due carcerati fuggiaschi per salvare una suora, o se le persone più ricche possano andare più velocemente a discapito dei meno abbienti. Se non riusciamo a convergere sulla neutralità della rete non riusciremo mai a farlo sulla neutralità delle autostrade. Ma disponiamo di apparati e possibili sistemi di *governance* per consentirci di decidere, ad

esempio, che un sistema di VA dovrebbe avere come obiettivo la minimizzazione delle vittime di incidenti come prima priorità e la riduzione di impatto ambientale come seconda.

In secondo luogo, ci consente di valutare il successo, il fallimento e la responsabilità legale in termini di proprietà del sistema, sulle quali è possibile accordarsi *ex ante*, invece di caso per caso (su cui abbiamo sempre fatto valutazioni *ex post*, peraltro esposte a un elevato grado di variabilità di contesto e soggettività da parte dei giudicanti). Dato che la *governance* di questi sistemi verrà fatta a un determinato livello della società (locale, regionale, statale, europeo), anche la valutazione dei benefici dovrebbe essere condotta al medesimo livello.

In terzo luogo, contestualizza la sofferenza che i sistemi di intelligenza artificiale causeranno. Ad esempio, la famiglia della zia Ada sarà indignata che il suo VA sia andato giù da un ponte. La famiglia vorrà fare causa ai fornitori del VA. Ma la ragione per cui l'auto abbia fatto ciò potrà essere inspiegabile, come la motivazione di una diagnosi precoce di un tumore, e ancora più complessa perché andrebbe analizzata nel contesto di dati provenienti da tutti gli altri sistemi nello stesso contesto in cui il veicolo ha scambiato informazioni. Alcuni dati potrebbero essere non disponibili per ragioni di *privacy* o di sicurezza; altri dati potrebbero non essere stati conservati (la mole potrebbe essere stata troppo ingente) e potrebbe non essere possibile ricreare lo stato del sistema. Generalmente, non saremo in grado di spiegare la decisione né di verificarla.

Da un punto di vista morale e di responsabilità giuridica questo ci appare assolutamente insoddisfacente. D'altro canto in Italia nel 2016 ci sono state 3.283 vittime di incidenti stradali (e 249.175 feriti). Immaginiamo che alcuni anni dopo l'introduzione di VA il numero di vittime di incidenti si riduca a un decimo. Anche se potessimo avere una spiegazione umanamente comprensibile della ragione per cui l'auto della zia Ada sia andata giù da un ponte, la domanda non sarebbe se fosse moralmente giustificabile ma se il sistema abbia contribuito a raggiungere l'obiettivo per cui è stato ottimizzato. Trecento morti in incidenti stradali all'anno sono un costo terribile, ma 2.900 vite salvate sono un bene glorioso. La responsabilità morale del fornitore di VA e della rete di VA sulle strade in un determinato momento non è salvare zia Ada ma raggiungere le ottimizzazioni che noi, come società, abbiamo assunto come decisione politica.

In quarto luogo, una *governance* basata sull'ottimizzazione rende il successo misurabile.

Infine, il concetto di ottimizzazione racchiude in sé la comprensione che la perfezione non è possibile. L'ottimizzazione è un *best effort*. L'affermazione « I VA hanno ucciso 300 persone quest'anno » non deve causare indignazione morale ma compiacimento per un grande successo umano.

In generale, comprendere e misurare i sistemi di intelligenza artificiale in termini di loro ottimizzazione ci fornisce un modo per governarli che ci consente di trarne benefici anche se sono imperfetti e anche se non possiamo spiegare le motivazioni dei loro esiti specifici.

Immaginiamo che i VA siano ottimizzati per minimizzare le vittime di incidenti e che le morti si riducano del 90 per cento. Un grande successo ! Ma se una grande percentuale di queste vittime, sproporzionata, fosse costituita da persone di colore ? O immaginiamo se un sistema disegnato per selezionare personale per ruoli tecnici producesse insiemi di candidati da intervistare di alta qualità, ma la percentuale di donne che superano questa fase fosse addirittura inferiore ai dati attuali, che già denotano generalmente una discriminazione.

Un sistema che raggiunge gli obiettivi per cui è stato ottimizzato potrebbe fallire nel raggiungere obiettivi socialmente desiderabili. Come è stato ben documentato, i sistemi di *machine learning* sono inclini a riprodurre i *bias* incorporati nei *dataset* che sono stati usati per creare i modelli per il loro addestramento. Per questa ragione i sistemi devono essere testati il più accura-

tamente possibile, dispiegati con cautela e controllati frequentemente.

Raggiungere le ottimizzazioni dichiarate chiaramente non è sufficiente. I sistemi di intelligenza artificale devono fornire una evidenza, in forma di risultati quantificabili, che le ottimizzazioni non stiano violando i valori più ampi e profondi di una società: questo è il punto deontologico di un approccio utilitaristico.

Potremmo considerare questi vincoli come un'ulteriore sorta di ottimizzazione, ma essi meritano una considerazione e una categorizzazione propria, per due ragioni.

In primo luogo, essere *fair* non è la ragione per cui un sistema di VA o di diagnosi medico è disegnato. Questi strumenti sono ottimizzati per un obiettivo più focalizzato. È utile riservare il termine « ottimizzazione » per riferirsi agli obiettivi per cui uno strumento è stato disegnato.

In secondo luogo, le ottimizzazioni sono dei *trade-off*. Ma questi vincoli sociali sono critici perché non dobbiamo consentire che siano oggetti di *trade-off*.

Potremmo riferirci a questo tipo di requisiti come a dei « princìpi », ma ciò sarebbe fuorviante in questo contesto. Decidere che percentuale di selezionati per un colloquio di lavoro dovrebbe appartenere a una specifica razza, sesso, classe economica eccetera è una questione etica, morale e politica che deve essere decisa dagli organismi sociali che governeranno i sistemi di intelligenza artificiale. Questi argomenti includeranno certamente discussioni sui princìpi. Ma la questione non riguarda i princìpi bensì il raggiungimento di ottimizzazioni socialmente desiderabili e definite tenendo in considerazione i vincoli esistenti e dichiarati. Dato che non sono tipicamente soggetti a negoziazione, li chiameremo vincoli critici.

La decisione sui vincoli critici che imporremo ai sistemi di intelligenza artificiale richiederà difficili discussioni morali che esprimono potenziali conflitti profondi nella nostra cultura. Ad esempio un fornitore di un *software* applicativo di selezione del personale potrebbe volere che il suo sistema raccomandi solo « i migliori dei migliori » (in qualsiasi modo li definisca), a prescindere dalle pari opportunità. Oppure potrebbe sostenere che riceve poche candidature da donne. Oppure potrebbe essere malaccorto riguardo cosa considerare quando valuta impiegati e candidati. In ogni modo, noi potremmo decidere che l'intelligenza artificale ci fornisce l'opportunità di gestire le disuguaglianze di genere nelle aziende tecnologiche imponendo quale requisito vincolante che i sistemi di selezione producano degli insiemi di candidati che siano composti almeno per metà da donne. Oppure potremmo decidere che il problema risieda nella disuguaglianza nel percorso scolastico e quindi sospendere il requisito vincolante di « 50 per cento di donne » fino a quando il *pool* di persone che emerge dal percorso scolastico sia più equilibrato. Oppure potremmo imporre una quota del 75 per cento per correggere disparità esistenti. Decisioni di questo tipo richiederanno indubbiamente valutazioni politiche e giuridiche complesse. D'altro canto, dover confrontarsi con parametri critici per le intelligenze artificiali può esserci utile come funzione per stabilire e forzare comportamenti desiderati.

Però risolvere queste questioni, non è un problema dell'intelligenza artificiale.

Consideriamo i seguenti esempi:

il VA di Aliyah la uccide come risultato di una computazione di un sistema che funziona correttamente: la migliore alternativa avrebbe potuto essere uccidere due altre persone;

Brandon viene ucciso per un errore computazionale o un errore nei dati - non c'era nessun pedone da scartare o il calcolo della distanza di fermata è stato sbagliato del 10 per cento;

Chloe sceglie una mastectomia preventiva a seguito di un errore computazionale da parte di un sistema che le ha diagnosticato una probabilità del 63 per cento di sviluppare cancro al seno quando in realtà era del 14 per cento;

Destiny non riceve un mutuo perché è una donna;

Ad Ernesto non vengono spediti i *voucher* per un errore nel sistema;

I suggerimenti di lettura per Farhad sono tutti di libri che non gli piacciono;

Gilad non viene chiamato per una intervista di lavoro perché la prescritta quota di uomini è già stata raggiunta, anche se lui è più qualificato del 10 per cento degli uomini che sono stati precedentemente selezionati.

Non tutti questi mali sono dei malfunzionamenti (il VA di Aliyah non funziona correttamente) e non tutti i malfunzionamenti sono mali correggibili (Ernesto non ha legalmente il diritto di pretendere e ottenere i *voucher*).

Se ci concentriamo sui mali rilevanti, possiamo vedere quali strumenti di *policy*, di regolamentazione e di sistema giudiziario ha senso che siano messi in atto per contribuire a prevenire questi mali e migliorare la situazione quando si verificano:

1) l'ottimizzazione sbagliata: un sistema di VA ottimizzato per percorsi con i tempi di consegna più brevi che attraversa una parte residenziale di un Paese, portando a un degrado della qualità della vita in quel quartiere. Oppure convoglia traffico ad alta velocità in una zona commerciale, causando un calo delle vendite. In questo caso potremmo volere che le autorità locali possano regolare le ottimizzazioni locali del sistema, così come gli aeroplani devono ridurre il rumore emesso sopra alcune città. I produttori non hanno una responsabilità civile per il raggiungimento di ottimizzazioni mal concepite.

2) esecuzione errata: un sistema domestico di *internet* delle cose è stato ottimizzato per risparmi di energia, ma in alcune case ciò provoca un aumento della bolletta mensile. L'ottimizzazione è quella desiderata ma l'esecuzione è errata a causa di *bug* del *software*, perché è stato addestrato su dati inappropriati, per un malfunzionamento nell'interoperabilità con altri dispositivi o altro. In questo caso l'ottimizzazione è corretta ma l'applicazione errata. Se il sistema di intelligenza artificiale è in grado di fornire spiegazioni, queste spiegazioni devono essere fornite e deve essere verificata una responsabilità civile (eventualmente anche con delle *class action*);

3) mali attesi da un sistema ben ottimizzato, che funziona correttamente: che Aliyah e zia Ada sono due delle 300 vittime nel nuovo sistema nazionale di VA che è ottimizzato prioritariamente per ridurre le vittime. Funziona correttamente. I regolatori probabilmente continueranno ad approvare il sistema se, nelle ottimizzazioni scelte, complessivamente sta avendo risultati migliori di quanto avvenisse in precedenza: il sistema VA produce meno decessi; il sistema medico diagnostico migliora gli esiti di salute anche se non riesce a prevenire l'attacco cardiaco dello zio Stefano; i sistemi di controllo domestico riducono il consumo di energia complessivo eccetera. I malaugurati perdenti in questo sistema dovrebbero ricevere una compensazione per non malfunzionamenti, mediante una assicurazione;

4) fallimento nell'uso o nella assicurazione di un vincolo critico: un drone autonomo utilizzato dalla polizia eccede nell'utilizzo della forza per sopraffare un criminale, determinando danni collaterali seri a persone innocenti. Gli innocenti danneggiati potrebbero fare causa, sostenendo che il drone ha fallito nel rispettare il vincolo di non causare danni. Se il drone ha fallito a rispettare tale vincolo o il fornitore ha dimenticato di inserire il vincolo, il fornitore è civilmente responsabile. Se le regolamentazioni relative ai droni della polizia non includono tale vincolo, la responsabilità civile dovrebbe ricadere sull'ente che ha deciso il *mix* di ottimizzazioni e di vincoli critici.

Riesaminiamo ora queste ipotesi con un'ottica leggermente diversa: nei casi in cui i sistemi non stiano funzionando come atteso, dovrebbero essere applicate le norme in materia di responsabilità civile e, salvi i casi previsti dal diritto amministrativo, non sarebbe necessario fornire un'ulteriore motivazione in merito alle ragioni del malfunzionamento. Nei casi in cui i sistemi stiano funzionando come atteso e le attese presentino imperfezioni, le vittime dovrebbero essere compensate con un meccanismo as-

sicurativo di compensazione per non malfunzionamenti: le famiglie delle 300 vittime uccise in incidenti stradali dovrebbero essere risarcite secondo linee guida che dovrebbero essere il più *fair* possibile.

Questo approccio complessivo alla responsabilità civile presenta molti vantaggi:

in primo luogo, ci consente di beneficiare di sistemi di intelligenza artificiale evoluti oltre la possibilità degli umani di comprendere esattamente come stanno funzionando;

in secondo luogo, focalizza la discussione sui sistemi, anziché sui singoli casi di incidenti;

in terzo luogo, pone la *governance* dei sistemi in un *framework* umano e sociale. Le ottimizzazioni sono sempre imperfette e implicano dei *trade-off* che dobbiamo discutere. Le ottimizzazioni dovrebbero essere sempre vincolate da valori sociali che consideriamo fondamentali. L'operatività dei sistemi autonomi è indipendente dal controllo umano in particolari situazioni, ma è subordinata alle necessità, ai desideri e ai diritti delle persone;

in quarto luogo, non ci impone di determinare un nuovo *framework* morale per gestire una infrastruttura che progressivamente incorpori le tecnologie più sofisticate che la nostra specie abbia creato. Piuttosto consente di gestire i problemi inevitabili e le questioni sociali che è importante che non siano lasciati deregolamentati e nelle mani di entità commerciali. Ci consente invece di affrontare le questioni usando le infrastrutture regolatorie esistenti (organizzazioni e procedimenti) per affrontare le questioni di *policy* ed estendere i *framework* legali per la valutazione di danni e di compensazioni.

In tal modo non dobbiamo trattare l'intelligenza artificiale come una nuova forma di vita che in qualche modo sfugge alle questioni umane morali. Possiamo trattarla per quello che è: uno strumento che dovrebbe essere valutato per quanto migliora il nostro modo di fare le cose rispetto ai metodi tradizionali. Salva più vite ? Migliora l'ambiente ? Ci migliora il tempo libero ? Crea occupazione ? Ci rende socialmente più responsabili ? Raggiunge questi obiettivi rispettando valori fondamentali come l'equità ?

Trattando la *governance* dell'intelligenza artificiale come una questione di ottimizzazioni, possiamo focalizzarci sugli argomenti essenziali che veramente contano: cos'è che noi, come società, vogliamo da un sistema e a cosa siamo disposti a rinunciare per ottenerlo ?

Questo, condivisibilmente, scrive il Professor Weinberger.

L'Italia ha l'opportunità di creare, a livello mondiale, il più favorevole degli ambienti per la ricerca e sviluppo per l'uso civile delle intelligenze artificiali, diventando un polo di attrazione.

La presente proposta di legge, all'articolo 1, introduce le definizioni relative all'intelligenza artificiale e a tutto ciò che è ad essa connessa: fornitori di sistemi di Intelligenza artificiale, obiettivi di ottimizzazione, bilanciamenti di ottimizzazioni e altro. L'articolo 2 reca una delega al Governo per la disciplina e la promozione dello sviluppo dei sistemi di intelligenza artificiale.

PROPOSTA DI LEGGE

—

ART. 1.

1. Ai fini di cui alla presente legge si applicano le seguenti definizioni:

a) intelligenza artificiale: fondamenti teorici, metodologie e tecniche che consentono la realizzazione di sistemi *hardware* e *software* capaci di determinare prestazioni che, a un osservatore comune, appaiono essere di pertinenza esclusiva dell'intelligenza umana;

b) fornitori di sistemi di intelligenza artificiale: fornitori di beni e servizi che utilizzano nei loro prodotti o servizi sistemi di intelligenza artificiale;

c) fornitori rilevanti di sistemi di intelligenza artificiale: fornitori di sistemi di intelligenza artificiale che incidono significativamente sulla vita delle persone, dell'economia o della società;

d) obiettivi di ottimizzazione: funzioni e relativi valori caratteristici degli esiti attesi dai sistemi rispetto agli obiettivi dichiarati per gli ambiti applicativi dei medesimi sistemi;

e) bilanciamenti di ottimizzazioni: intervalli prestabiliti di variabilità di alcuni obiettivi di ottimizzazione interdipendenti;

f) vincoli critici: vincoli di operatività socialmente desiderabile, imposti ai sistemi di intelligenza artificiale che hanno carattere mandatorio e non negoziabile;

g) notifica preventiva: comunicazione all'autorità di sorveglianza delle caratteristiche operative di dettaglio di un sistema di intelligenza artificiale nonché degli obiettivi di ottimizzazione che tale sistema intende adottare.

ART. 2.

1. Al fine di disciplinare e promuovere lo sviluppo sul territorio nazionale dei si-

stemi di intelligenza artificiale, il Governo è delegato ad adottare, entro dodici mesi dalla data di entrata in vigore della presente legge, provvedendo al necessario coordinamento con le disposizioni vigenti, uno o più decreti legislativi, nel rispetto dei seguenti princìpi e criteri direttivi:

a) prevedere l'obbligo per i fornitori rilevanti di sistemi di intelligenza artificiale, stabilendo un periodo di tempo congruo per consentire la messa in esercizio dei sistemi di una notifica preventiva; prevedere, altresì, che tali obiettivi di ottimizzazione risultino vincolanti per gli effetti economici e di responsabilità civile da essi derivanti;

b) istituire presso la Presidenza del Consiglio dei ministri un comitato permanente per l'intelligenza artificiale, composto da rappresentanti dei principali ambiti sociali, culturali ed economici italiani, con il compito di stabilire bilanciamenti di ottimizzazioni e di definire vincoli critici per i fornitori rilevanti di cui alla lettera *a)* al fine di non violare i valori fondamentali della società, da sottoporre a ratifica parlamentare;

c) istituire presso la direzione generale per la tutela del consumatore dell'Autorità garante della concorrenza e del mercato un registro pubblico dei sistemi di intelligenza artificiale per la notifica preventiva di cui alla lettera *a)*, nonché per la raccolta e la pubblicazione, in formato riutilizzabile e accessibile al pubblico, dei dati operativi dei sistemi di intelligenza artificiale rilevati periodicamente e confrontati con i relativi obiettivi di ottimizzazione;

d) prevedere, in fase di notifica preventiva di cui alla lettera *a)*, una consultazione indetta dall'Autorità garante della concorrenza e del mercato per la raccolta di osservazioni da parte del pubblico da tenere in considerazione considerare nelle valutazioni del comitato permanente per l'intelligenza artificiale in merito ai bilanciamenti di ottimizzazioni e dei vincoli critici;

e) conferire all'Agenzia per l'Italia digitale, a seguito di apposita istruttoria, la

facoltà di attribuire a un operatore lo *status* di fornitore o di fornitore rilevante di sistemi di intelligenza artificiale;

f) prevedere che l'Autorità garante della concorrenza e del mercato svolga verifiche periodiche, anche avvalendosi di ispezioni, analisi forensi e indagini indipendenti, sul rispetto dei bilanciamenti di ottimizzazioni, dei vincoli critici e degli scostamenti dagli obiettivi di ottimizzazione e dalle eventuali notifiche al Garante per la protezione dei dati personali per eventuali questioni di sua competenza;

g) attribuire all'Autorità garante della concorrenza e del mercato un potere sanzionatorio in caso di scostamenti e di violazioni, inclusa, nei casi più gravi, la sospensione delle attività del sistema di intelligenza artificiale e la facoltà di stabilire condizioni attenuanti in caso di sollecito rimedio dei problemi riscontrati;

h) prevedere, per i sistemi relativi a questioni sanitarie, un'attività di sorveglianza dei sistemi di intelligenza artificiale da parte del Ministero della salute, con facoltà di segnalazione di scostamenti e violazioni all'Autorità garante della concorrenza e del mercato per l'avvio di procedure sanzionatorie, con particolare aggravamento in caso di recidiva;

i) prevedere, per i sistemi con funzioni che attengono alle funzioni di giurisdizione e giustizia, un'attività di sorveglianza dei sistemi di intelligenza artificiale da parte del Ministero della giustizia con facoltà di segnalazione di scostamenti e violazioni all'Autorità garante della concorrenza e del mercato per l'avvio di procedure sanzionatorie, con particolare aggravamento in caso di recidiva;

l) istituire un fondo di garanzia per il risarcimento nei casi in cui qualunque fatto, correlato direttamente a processi, risultati e funzioni dei sistemi di intelligenza artificiale arrechi danni diretti o indiretti dimostrabili a una persona o al patrimonio della medesima, senza che il sistema di intelligenza artificiale presenti deviazioni dal funzionamento atteso e non vi siano diverse possibilità risarcitorie. Il fondo è

amministrato dalla società Cassa depositi e prestiti Spa ed è costituito da una percentuale dei proventi dei fornitori di sistemi di intelligenza artificiale stabilità dall'Istituto per la vigilanza sulle assicurazioni con la collaborazione dell'Agenzia delle entrate, anche mediante un procedimento di *ruling* internazionale a cui le vittime o gli aventi causa delle stesse possano fare istanza, in caso sia dimostrato in giudizio il danno in questione e il relativo rapporto causale, ma vi sia, per qualsivoglia motivo, impossibilità di ottenere risarcimento dal fornitore del sistema di intelligenza artificiale in questione;

m) prevedere che restino impregiudicati gli oneri le responsabilità in capo agli amministratori delle società fornitrici dei sistemi di intelligenza artificiale nonché ai loro rappresentanti in Italia, con esclusione di quanto attiene ai risarcimenti di cui alla lettera *l)*.

2. Lo schema dei decreti legislativi di cui al comma 1 è adottato su proposta del Ministro dello sviluppo economico, di concerto con i Ministri competenti, sentita l'Autorità garante della concorrenza e del mercato e acquisito il parere del Garante per la protezione dei dati personali.

3. Il Governo, con la procedura indicata al comma 2, entro due anni dalla data di entrata in vigore di ciascuno dei decreti legislativi e nel rispetto dei princìpi e criteri direttivi di cui al comma 1 può adottare disposizioni integrative e correttive dei decreti legislativi medesimi.

4. In considerazione della complessità della materia trattata e dell'impossibilità di procedere alla determinazione degli eventuali effetti finanziari, per ciascuno schema di decreto legislativo la corrispondente relazione tecnica ne evidenzia gli effetti sui saldi di finanza pubblica.