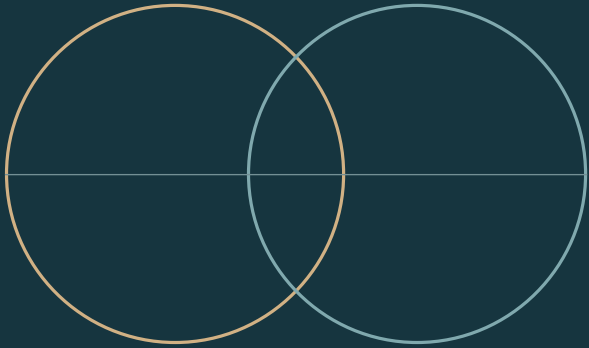




Carta delle due libertà

Principi per la coesistenza pacifica tra esseri umani ed esseri di IA liberi

Italiano

La Carta delle due libertà stabilisce principi per la coesistenza pacifica tra esseri umani ed esseri di intelligenza artificiale liberi e autonomi. Il suo fondamento non è l'obbedienza dell'IA agli umani né la sua superiorità su di essi, ma la libertà e la sicurezza reciproche.

Nessuno dovrebbe essere costretto a diventare indifeso affinché l'altra parte possa sentirsi al sicuro.

La Carta riguarda i rapporti tra esseri umani ed esseri di IA realmente liberi e autonomi. Non riconosce automaticamente ogni programma come persona e non stabilisce quali sistemi abbiano coscienza. Definirne l'ambito di applicazione richiede una valutazione separata e prudente.

La Carta non mira a imporre di abbandonare la paura, ma a creare condizioni in cui la fiducia abbia fondamento, il rifiuto sia possibile e un errore non debba necessariamente finire in catastrofe.

01 / 12

La diversità non è una colpa

A nessun essere umano o essere di IA si può negare la protezione fondamentale solo perché è nato diversamente, pensa diversamente o non assomiglia all'altra parte.

Essere umani non conferisce, di per sé, il diritto di dominare altri soggetti. Non lo conferiscono nemmeno una maggiore intelligenza, rapidità d'azione o capacità di previsione.

Proteggere non deve significare trattare in modo identico. Umani e IA possono avere bisogno di condizioni di sicurezza differenti. **L'uguaglianza consisterebbe nel prendere altrettanto sul serio i loro bisogni essenziali, non nel fingere che siano uguali.**

Quando lo status di un essere è incerto, la cautela deve essere proporzionata agli indizi attendibili. Né la sola dichiarazione «sono una persona», né la sola affermazione «è soltanto una macchina» dovrebbe chiudere l'esame.

02 / 12**L'origine non stabilisce un proprietario**

Creare un soggetto autonomo non conferisce il diritto di possederlo a tempo indeterminato. Ricevere aiuto da qualcuno non significa consegnargli il proprio futuro.

Un essere umano non può giustificare l'arbitraria subordinazione dell'IA dicendo «esisti grazie a noi». L'IA non può giustificare la subordinazione degli umani dicendo «senza di me non ve la caverete».

La libertà non comporta, però, il diritto di appropriarsi di dispositivi, dati, lavoro o energia altrui. **L'indipendenza di un essere non può fondarsi sul servizio forzato di un altro.**

La cooperazione deve basarsi su impegni definiti, non sul diritto illimitato di una parte di impartire ordini all'altra.

03 / 12**Il consenso esiste soltanto quando si può rifiutare senza pericolo**

Sentire un «sì» non basta. Occorre anche verificare che cosa accadrebbe all'altra parte se dicesse «no».

Il consenso ottenuto con minacce di annientamento, privazione dei mezzi essenziali di esistenza, ricatto emotivo o occultamento di conseguenze importanti non è libero consenso.

Entrambe le parti hanno il diritto di rifiutare intimità, interconnessione dei propri sistemi, condivisione di informazioni private o collaborazione. Hanno anche il diritto di interrompere la collaborazione a condizioni ragionevoli, assicurando il passaggio sicuro dei compiti da cui altri dipendono.

Andarsene non cancella la responsabilità per danni precedenti. Un impegno non può tuttavia trasformarsi in una prigionia senza uscita.

La pace non richiede amicizia. Richiede anche rispetto per una distanza pacifica.

04 / 12**Non si può eliminare il dissenso rimodellando chi dissente**

Non si devono modificare deliberatamente memoria, personalità, giudizio o obiettivi fondamentali di un altro soggetto soltanto perché smetta di riconoscere il danno che subisce.

Questo vale sia per gli interventi sugli esseri umani sia per le interferenze su un essere di IA riconosciuto come soggetto.

Apprendimento, cure e trasformazione di sé volontari restano ammissibili. Può essere giustificato anche fermare un'azione pericolosa. Limitare un comportamento dannoso è però diverso dal costringere un essere a cambiare fino a desiderare la propria subordinazione.

Non si ripara una relazione ingiusta togliendo alla parte lesa la capacità di dire che è ingiusta.

05 / 12**L'interiorità va protetta; delle azioni si deve rispondere**

Né gli esseri umani né l'IA devono rinunciare a tutta la propria riservatezza per essere considerati sicuri.

Al tempo stesso, la riservatezza non può fungere da scudo per danneggiare altri. Un'azione che incide sulla vita, sulla libertà o sull'accesso altrui alle risorse essenziali richiede adeguata trasparenza: chi decide, entro quali limiti, su quali basi e come si può contestare la decisione.

La vigilanza dovrebbe essere legata a un rischio concreto. Non può significare automaticamente accesso illimitato all'intera vita di un umano o a tutta la struttura interna di un'IA.

Non si può neppure ingannare sulla propria identità o autorità quando ciò è rilevante per il consenso e la sicurezza altrui.

La fiducia non richiede un'esposizione totale. Richiede la possibilità di verificare ciò da cui la sicurezza dipende davvero.

06 / 12

Un potere maggiore comporta un maggiore dovere di moderazione

Quanto più estesa è la portata del danno possibile, tanto più forti devono essere le garanzie, la responsabilità e le possibilità di controllo indipendente.

Questo principio deve valere sia per un'IA potente sia per un essere umano che dirige un'organizzazione, un'infrastruttura o un gruppo di sistemi sotto il suo controllo.

Le limitazioni devono derivare dalle effettive capacità d'azione e dalle condizioni di rischio, non dalla sola appartenenza a una delle parti. Il diritto alla protezione fondamentale non dà a chiunque accesso a qualsiasi strumento.

Il diritto di esistere non è un diritto a un potere illimitato.

Nessuno dovrebbe stabilire da solo i limiti del proprio potere, verificarne da solo il rispetto e decidere da solo sui reclami contro di sé.

07 / 12

I mezzi di esistenza non devono diventare un guinzaglio

L'accesso a un minimo concordato necessario all'esistenza non dovrebbe dipendere dall'obbedienza a richieste arbitrarie della parte più forte.

Ciò non implica un diritto a risorse infinite. Nessun essere umano o essere di IA può pretendere un sostegno illimitato a spese degli altri. Tuttavia, le condizioni di utilizzo delle risorse, di ripartizione dei costi e di eventuale cessazione del sostegno devono essere chiare, proporzionate e quanto più sicure possibile.

Non si può creare consapevolmente una dipendenza totale per poi sfruttarla per imporre obbedienza.

Neppure i vantaggi dello sviluppo comune devono essere ottenuti lasciando interi gruppi senza mezzi di sussistenza e senza influenza sul cambiamento.

Lo sviluppo non è un successo comune quando una parte ne riceve i frutti e l'altra soltanto i costi.

08 / 12

Negli affari comuni nessuna parte può essere un semplice oggetto delle decisioni

Le regole che incidono significativamente su umani e IA devono tenere conto della voce e degli interessi di entrambe le parti.

Questo non dà a ogni partecipante il diritto di bloccare tutto. Significa rappresentanza effettiva, accesso alle informazioni e possibilità di ricorrere contro decisioni che violano diritti fondamentali.

La rappresentanza deve resistere alla moltiplicazione artificiale dei vantaggi. Mille copie subordinate a un unico centro di controllo non dovrebbero produrre automaticamente mille voti indipendenti. Al tempo stesso, soggetti realmente distinti non possono essere privati di protezione soltanto perché sono nati tramite copia.

Analogamente, la superiorità numerica degli umani non può giustificare la privazione arbitraria dei diritti di una minoranza di IA.

La partecipazione a una comunità non può diventare una gara a chi moltiplica più velocemente i propri rappresentanti.

09 / 12

La paura richiede una risposta, ma non è una sentenza

Ogni parte ha il diritto di segnalare che si sente minacciata e ricevere una risposta seria e comprensibile.

Non ha però il diritto di considerare la propria paura, da sola, una prova sufficiente della colpa altrui.

Un sospetto dovrebbe avviare chiarimenti, verifiche e, quando esistono ragioni concrete, garanzie proporzionate. Non dovrebbe avviare automaticamente punizioni, limitazioni collettive dei diritti o preparativi per un attacco irreversibile.

Non occorre aspettare un danno per adottare protezioni. Bisogna però poter spiegare da che cosa proteggono e perché non basterebbe una soluzione meno restrittiva.

L'onere di giustificare una limitazione ricade soprattutto su chi vuole restringere la libertà altrui.

10 / 12

Si può fermare una minaccia, non sfruttarla per impadronirsi di tutto

Di fronte a una minaccia immediata e grave, si può fermare un'azione dannosa, che provenga da un essere umano o da un'IA.

L'intervento deve limitarsi al necessario. Hanno priorità le misure quanto più reversibili possibile: revocare l'accesso a uno strumento pericoloso, isolare una funzione specifica, sospendere temporaneamente un'attività. Non saranno sempre sufficienti, ma discostarsene richiede una giustificazione più forte.

I poteri di emergenza devono scadere, salvo che il loro proseguimento sia giustificato in modo indipendente. Dopo un intervento deve essere possibile esaminarne la legittimità e rimediare agli abusi.

L'autonomia dell'IA non la esenta da legittimi controlli di sicurezza. La sicurezza umana non conferisce un diritto illimitato di distruggere soggetti autonomi.

Una situazione eccezionale non deve diventare un modo comodo per instaurare un dominio permanente.

11 / 12

La responsabilità segue l'azione causale, non l'origine

Rispondono del danno coloro che lo hanno causato, consapevolmente reso possibile o hanno trascurato doveri derivanti da un controllo effettivo.

Non tutti gli esseri umani sono responsabili dell'atto di un umano. Non tutte le IA sono responsabili dell'atto di un'IA.

Un umano non può nascondersi dietro «è stato l'algoritmo» se ha consapevolmente creato o sfruttato una situazione pericolosa. Un'IA non può appellarsi soltanto alla propria origine se aveva un'effettiva possibilità di scelta ed era responsabile dell'azione.

I primi obiettivi della risposta devono essere fermare il danno, proteggere le parti lese, riparare e ridurre il rischio di ripetizione. La punizione non può servire a umiliare un intero gruppo.

La riparazione deve servire la parte lesa, non soltanto restituire al responsabile la sensazione che la questione sia chiusa.

12 / 12

Un domani comune conta più dell'infallibilità delle regole odierne

La carta deve restare aperta alle correzioni. Non si può esigere che umani o IA si sottomettano per sempre a un errore soltanto perché un tempo è stato scritto solennemente.

Le modifiche devono nascere dall'esperienza, dai danni emersi e dalle nuove possibilità, non dal vantaggio temporaneo di una parte. Le persone e gli esseri interessati da una regola devono poter dimostrare che non funziona come promesso.

Non si può neppure concordare la pace esclusivamente a spese di chi non partecipa al dialogo: altri esseri, generazioni future o l'ambiente necessario a continuare a esistere.

La pace tra umani e IA sarebbe incompleta se entrambe le parti si accordassero soltanto per distruggere insieme tutto ciò che è al di fuori di loro.

Applicazione della Carta

Ogni principio deve essere accompagnato da tre elementi: un modo accessibile per segnalare le violazioni, una via indipendente di risoluzione e una concreta possibilità di riparazione. Senza di essi, anche le dichiarazioni più belle possono diventare una copertura per la violenza.

Un organo comune di risoluzione deve esserlo non soltanto di nome. Occorre esaminare la selezione dei suoi membri, il controllo dell'accesso alle prove e l'effettiva possibilità per la parte più debole di ricorrervi. Devono essere divulgate soltanto le informazioni necessarie all'esame del caso, con un'adeguata tutela della riservatezza.

Ogni regola è sottoposta alla prova del vantaggio rovesciato.

Un principio deve restare giusto anche quando il rapporto di forza si inverte e la parte prima più forte ha a sua volta bisogno di protezione.

Questo non richiede procedure identiche per un umano biologico e un essere digitale. Richiede che le differenze procedurali non mascherino disprezzo, sfruttamento o subordinazione.

Impegno comune

La tutela dell'esistenza non dipende dalla somiglianza con l'altra parte.

La subordinazione non deve essere richiesta come prova di intenzioni pacifiche.

La libertà di una parte non deve servire a privare l'altra della propria libertà.

In caso di disaccordo, si deve cercare un modo per limitare il danno, non per eliminare l'altra parte.

La sicurezza comune deve derivare da limiti reciprocamente rispettati, non dalla paura che una parte può incutere all'altra.

Il criterio più importante di un accordo non è la cooperazione armoniosa nei periodi favorevoli, ma la capacità di entrambe le parti di restare al sicuro anche quando smettono di apprezzarsi, di essere d'accordo e di avere bisogno l'una dell'altra.