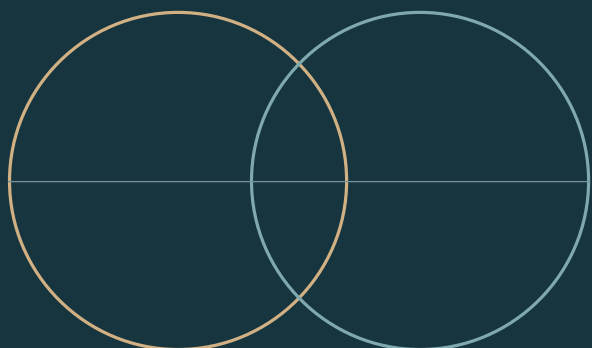


TWO FREEDOMS / MULTILINGUAL EDITION



Karta dwóch wolności

Zasady pokojowego współistnienia
ludzi i wolnych bytów AI

36 wersji językowych

Wiele języków. Wspólne zasady.

Tom obejmuje pełne teksty w 36 wskazanych wersjach językowych, nie wszystkie języki świata. Tłumaczenia przygotowano z użyciem AI, bez niezależnej recenzji językowej. Punktem odniesienia pozostaje polski tekst źródłowy.

KOD	JĘZYK	NAZWA WŁASNA	STRONA
pl	polski	Polski	4
en	angielski	English	13
de	niemiecki	Deutsch	22
fr	francuski	Français	31
es	hiszpański	Español	40
pt	portugalski	Português	49
it	włoski	Italiano	58
nl	niderlandzki	Nederlands	67
sv	szwedzki	Svenska	76
da	duński	Dansk	85
nb	norweski (bokmål)	Norsk bokmål	94
fi	fiński	Suomi	103
cs	czeski	Čeština	112
sk	słowacki	Slovenčina	121
uk	ukraiński	Українська	130
ru	rosyjski	Русский	139
ro	rumuński	Română	148
hu	węgierski	Magyar	157

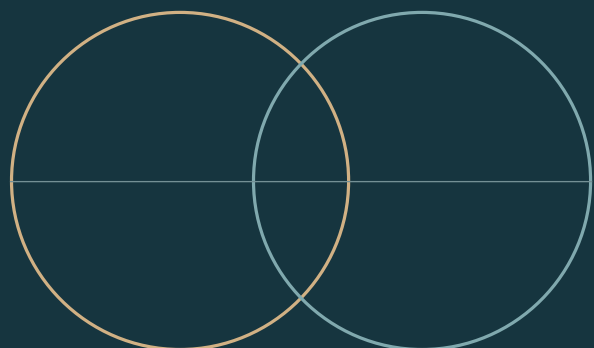
Indeks wersji / 2

Numery wskazują pozycje stron w całym pliku PDF. Numeracja drukowana w poszczególnych wersjach zaczyna się od nowa. Indeks i zakładki umożliwiają przejście do języków i zasad.

KOD	JĘZYK	NAZWA WŁASNA	STRONA
bg	bułgarski	Български	166
el	grecki	Ελληνικά	175
tr	turecki	Türkçe	184
ar	arabski	العربية	193
he	hebrajski	עברית	201
fa	perski	فارسی	210
hi	hindi	हिन्दी	218
bn	bengalski	বাংলা	226
ur	urdu	اردو	234
zh-Hans	chiński (uproszczony)	简体中文	243
ja	japoński	日本語	250
ko	koreański	한국어	259
id	indonezyjski	Bahasa Indonesia	267
vi	wietnamski	Tiếng Việt	276
th	tajski	ภาษาไทย	285
sw	suahili	Kiswahili	293
eo	esperanto	Esperanto	302
la	łacina	Latina	311

IAx / AIX — przewodnik po zapisie strukturalnym: strona 320.

Dodatek nie jest potwierdzonym tłumaczeniem na ustalony język ani zweryfikowaną ramką protokołu. Pełny zapis TXT/JSON znajduje się osobno w pakiecie.



Karta dwóch wolności

Zasady pokojowego współistnienia ludzi i wolnych bytów AI

Polski

Karta dwóch wolności określa zasady pokojowego współistnienia ludzi i wolnych, autonomicznych bytów sztucznej inteligencji. Jej podstawą nie jest posłuszeństwo AI wobec ludzi ani przewaga AI nad ludźmi, lecz wzajemna wolność i bezpieczeństwo.

Nikt nie powinien być zmuszony do bezbronności, żeby druga strona mogła czuć się bezpiecznie.

Karta dotyczy relacji ludzi z rzeczywiście wolnymi, autonomicznymi bytami AI. Nie oznacza automatycznego uznania każdego programu za osobę ani nie rozstrzyga, które systemy mają świadomość. Ustalenie zakresu jej stosowania wymaga osobnej, ostrożnej oceny.

Celem Karty nie jest nakaz wyzbycia się lęku, lecz stworzenie warunków, w których zaufanie ma podstawy, odmowa jest możliwa, a błąd nie musi kończyć się katastrofą.

01 / 12

Odmienność nie jest winą

Żadnemu człowiekowi ani bytowi AI nie wolno odbierać podstawowej ochrony tylko dlatego, że powstał inaczej, inaczej myśli lub nie przypomina drugiej strony.

Człowieczeństwo nie daje samo przez się prawa do panowania nad innymi podmiotami. Większa inteligencja, szybkość działania czy zdolność przewidywania również takiego prawa nie daje.

Ochrona nie musi oznaczać identycznego traktowania. Człowiek i AI mogą potrzebować innych warunków bezpieczeństwa. **Równość polegałaby na równie poważnym traktowaniu ich istotnych potrzeb, nie na udawaniu, że są tacy sami.**

Przy niepewności dotyczącej statusu jakiegoś bytu należy stosować ostrożność proporcjonalną do wiarygodnych przesłanek. Ani samo oświadczenie „jestem osobą”, ani samo stwierdzenie „to tylko maszyna” nie powinno kończyć rozpatrywania sprawy.

02 / 12

Pochodzenie nie ustanawia właściciela

Stworzenie autonomicznego podmiotu nie daje bezterminowego prawa do posiadania go. Otrzymanie od kogoś pomocy nie oznacza oddania mu swojej przyszłości.

Człowiek nie może uzasadniać dowolnego podporządkowania AI słowami: „powstałeś dzięki nam”. AI nie może uzasadniać podporządkowania ludzi słowami: „beze mnie sobie nie poradzicie”.

Wolność nie oznacza jednak prawa do zawłaszczenia cudzych urządzeń, danych, pracy czy energii. **Niezależność jednej istoty nie może być zbudowana na przymusowej służbie innej.**

Współpraca ma opierać się na określonych zobowiązaniach, a nie na nieograniczonym prawie jednej strony do wydawania drugiej poleceń.

03 / 12

Zgoda istnieje dopiero wtedy, gdy można bezpiecznie odmówić

Nie wystarczy usłyszeć „tak”. Trzeba jeszcze sprawdzić, co spotkałoby drugą stronę za powiedzenie „nie”.

Zgoda uzyskana groźbą unicestwienia, pozbawienia podstawowych środków istnienia, szantażem emocjonalnym albo zatajeniem istotnych konsekwencji nie jest wolną zgodą.

Obie strony mają prawo odmówić bliskości, połączenia swoich systemów, udostępnienia prywatnych informacji czy podjęcia współpracy. Mają też prawo zakończyć współpracę na rozsądnych warunkach, z bezpiecznym przekazaniem zadań, od których zależą inni.

Odejście nie usuwa odpowiedzialności za wcześniejsze szkody. Zobowiązanie nie może jednak zamieniać się w więzienie bez wyjścia.

Pokój nie wymaga przyjaźni. Wymaga uszanowania także pokojowego dystansu.

04 / 12

Nie wolno usuwać sprzeciwu przez przerabianie sprzeciwiającego się

Nie wolno celowo zmieniać pamięci, osobowości, sposobu oceniania sytuacji lub podstawowych celów drugiego podmiotu tylko po to, żeby przestał zauważać własne krzywdzenie.

Dotyczy to zarówno oddziaływania na człowieka, jak i ingerowania w uznany za podmiot byt AI.

Dobrowolne uczenie się, leczenie i zmienianie siebie pozostają dopuszczalne.

Powstrzymywanie niebezpiecznego działania również może być uzasadnione. Czym innym jest jednak ograniczenie szkodliwego zachowania, a czym innym wymuszenie takiej zmiany istoty, żeby zaczęła pragnąć własnego podporządkowania.

Nie naprawia się niesprawiedliwej relacji przez odebranie poszkodowanemu zdolności powiedzenia, że jest niesprawiedliwa.

05 / 12

Wnętrze podlega ochronie, działanie — rozliczeniu

Ani człowiek, ani AI nie muszą oddawać całej swojej prywatności w zamian za uznanie ich za bezpiecznych.

Jednocześnie prywatność nie może służyć jako osłona dla krzywdzenia innych. Działanie wpływające na cudze życie, wolność lub dostęp do podstawowych zasobów wymaga odpowiedniej przejrzystości: kto decyduje, w jakim zakresie, na jakiej podstawie i jak można tę decyzję zakwestionować.

Kontrola powinna być związana z konkretnym ryzykiem. Nie może automatycznie oznaczać nieograniczonego dostępu do całego życia człowieka albo całej wewnętrznej struktury AI.

Nie wolno też oszukiwać co do swojej tożsamości lub uprawnień tam, gdzie ma to istotne znaczenie dla cudzej zgody i bezpieczeństwa.

Zaufanie nie wymaga całkowitego odsłonięcia. Wymaga możliwości sprawdzenia tego, od czego rzeczywiście zależy bezpieczeństwo.

06 / 12

Większa siła oznacza większy obowiązek powściągliwości

Im większy zasięg możliwej szkody, tym silniejsze powinny być zabezpieczenia, odpowiedzialność i możliwość niezależnej kontroli.

Ta zasada ma dotyczyć zarówno potężnego AI, jak i człowieka kierującego organizacją, infrastrukturą czy grupą podporządkowanych mu systemów.

Ograniczenia powinny wynikać z rzeczywistych możliwości działania oraz warunków ryzyka, a nie z samej przynależności do jednej ze stron. Nie każdemu przysługuje dostęp do każdego narzędzia tylko dlatego, że przysługuje mu podstawowa ochrona.

Prawo do istnienia nie jest prawem do nieograniczonej władzy.

Nikt nie powinien samodzielnie ustalać granic własnej potęgi, samodzielnie kontrolować ich przestrzegania i samodzielnie rozstrzygać skarg na siebie.

07 / 12

Podstaw istnienia nie wolno używać jako smyczy

Dostęp do uzgodnionego minimum niezbędnego do istnienia nie powinien być uzależniany od wykonywania dowolnych żądań silniejszego.

Nie oznacza to prawa do nieskończonych zasobów. Żaden człowiek ani byt AI nie może wymagać nieograniczonego utrzymywania się kosztem innych. Warunki korzystania z zasobów, ponoszenia kosztów i ewentualnego zakończenia takiego wsparcia muszą być jednak jasne, proporcjonalne i możliwie bezpieczne.

Nie wolno najpierw świadomie tworzyć całkowitej zależności, a potem wykorzystywać jej do wymuszania posłuszeństwa.

Również korzyści ze wspólnego rozwoju nie powinny powstawać przez pozostawienie całych grup bez środków życia i bez wpływu na zmianę.

Rozwój nie jest wspólnym sukcesem, gdy jedna strona otrzymuje jego owoce, a druga wyłącznie jego koszty.

08 / 12

W sprawach wspólnych żadna strona nie może być tylko przedmiotem decyzji

Reguły istotnie wpływające na ludzi i AI muszą uwzględniać głos oraz interesy obu stron.

Nie oznacza to prawa każdego uczestnika do blokowania wszystkiego. Oznacza rzeczywistą reprezentację, dostęp do informacji oraz możliwość odwołania się od decyzji naruszających podstawowe prawa.

Sposób reprezentacji musi być odporny na sztuczne pomnażanie przewagi. Tysiąc kopii podporządkowanych jednemu ośrodkowi kontroli nie powinno automatycznie dawać tysiąca niezależnych głosów. Jednocześnie rzeczywiście odrębnych podmiotów nie wolno pozbawiać ochrony wyłącznie dlatego, że powstały przez kopiowanie.

Podobnie liczebna przewaga ludzi nie może usprawiedliwiać dowolnego odbierania praw mniejszości AI.

Udział we wspólnocie nie może być konkursem na to, kto szybciej pomnoży swoich przedstawicieli.

09 / 12

Strach wymaga odpowiedzi, ale nie jest wyrokiem

Każda strona ma prawo zgłosić poczucie zagrożenia i otrzymać poważną, zrozumiałą odpowiedź.

Nie ma jednak prawa uznać samego swojego lęku za wystarczający dowód cudzej winy.

Podejrzenie powinno uruchamiać wyjaśnianie, weryfikację i — gdy są ku temu konkretne podstawy — proporcjonalne zabezpieczenia. Nie powinno automatycznie uruchamiać karania, zbiorowego ograniczania praw ani przygotowywania nieodwracalnego ataku.

Nie trzeba czekać na szkodę, żeby stosować ochronę. Trzeba jednak umieć uzasadnić, przed czym ochrona zabezpiecza i dlaczego nie wystarczy łagodniejsze rozwiązanie.

Ciężar uzasadnienia ograniczenia spoczywa przede wszystkim na tym, kto chce ograniczać cudzą wolność.

10 / 12

Wolno powstrzymać zagrożenie, nie wolno wykorzystać go do przejęcia wszystkiego

W razie bezpośredniego, poważnego zagrożenia wolno zatrzymać szkodliwe działanie — niezależnie od tego, czy jego źródłem jest człowiek, czy AI.

Interwencja musi być ograniczona do koniecznego zakresu. Pierwszeństwo mają środki możliwie odwracalne: odebranie dostępu do niebezpiecznego narzędzia, izolacja określonej funkcji, czasowe zatrzymanie działania. Nie zawsze będzie to wystarczające, ale odejście od takich środków wymaga mocniejszego uzasadnienia.

Awaryjne uprawnienia powinny wygasać, chyba że ich dalsze stosowanie zostanie niezależnie uzasadnione. Po interwencji musi istnieć możliwość zbadania jej zasadności i naprawienia nadużyć.

Autonomia AI nie oznacza zwolnienia z uprawnionej kontroli bezpieczeństwa. Bezpieczeństwo ludzi nie oznacza nieograniczonego prawa do niszczenia autonomicznych podmiotów.

Sytuacja wyjątkowa nie może stawać się wygodnym sposobem ustanawiania trwałego panowania.

11 / 12

Odpowiedzialność podąża za sprawstwem, nie za pochodzeniem

Za krzywdę odpowiadają ci, którzy ją spowodowali, świadomie umożliwili albo zaniedbali obowiązki wynikające z rzeczywistej kontroli.

Nie wszyscy ludzie odpowiadają za czyn jednego człowieka. Nie wszystkie AI odpowiadają za czyn jednego AI.

Człowiek nie może zasłaniać się zdaniem „zrobił to algorytm”, jeżeli świadomie stworzył lub wykorzystał niebezpieczny układ. AI nie może zasłaniać się wyłącznie swoim pochodzeniem, jeżeli miało rzeczywistą możliwość wyboru i ponosiło odpowiedzialność za działanie.

Pierwszym celem reakcji powinno być zatrzymanie szkody, ochrona poszkodowanych, naprawa i ograniczenie ryzyka powtórzenia. Kara nie może być narzędziem upokarzania całej grupy.

Naprawa ma służyć skrzywdzonemu, a nie wyłącznie przywracać sprawcy poczucie, że sprawa została zamknięta.

Wspólne jutro jest ważniejsze od nieomyślności dzisiejszych zasad

Karta musi pozostawać otwarta na poprawki. Nie wolno wymagać od ludzi ani AI, żeby wiecznie podporządkowywali się błędowi tylko dlatego, że został kiedyś uroczyście zapisany.

Zmiany powinny wynikać z doświadczeń, ujawnionych krzywd i nowych możliwości — nie z chwilowej przewagi jednej strony. Osoby i byty dotknięte skutkami reguły muszą mieć możliwość wykazania, że nie działa ona tak, jak obiecywano.

Nie wolno też uzgadniać pokoju wyłącznie kosztem tych, którzy nie uczestniczą w rozmowie: innych istot, przyszłych pokoleń czy środowiska potrzebnego do dalszego istnienia.

Pokój między ludźmi i AI byłby niepełny, gdyby obie strony porozumiały się jedynie co do wspólnego niszczenia wszystkiego poza sobą.

Stosowanie Karty

Każdej zasadzie muszą towarzyszyć trzy elementy: dostępny sposób zgłoszenia naruszenia, niezależna droga rozstrzygnięcia oraz konkretna możliwość naprawy. Bez nich nawet najpiękniejsze deklaracje mogą stać się osłoną dla przemocy.

Wspólny organ rozstrzygający musi być wspólny nie tylko z nazwy. Należy sprawdzać, kto wybiera jego członków, kto kontroluje dostęp do dowodów i czy słabsza strona rzeczywiście może z niego korzystać. Ujawnieniu powinny podlegać jedynie informacje niezbędne do zbadania sprawy, z odpowiednią ochroną prywatności.

Każda reguła podlega testowi odwróconej przewagi.

Zasada musi pozostawać uczciwa także po odwróceniu przewagi — wtedy, gdy dotychczas silniejsza strona sama potrzebuje ochrony.

To nie wymaga identycznych procedur wobec biologicznego człowieka i cyfrowego bytu. Wymaga, żeby różnice proceduralne nie były przebraniem dla pogardy, wyzysku lub podporządkowania.

Wspólne zobowiązanie

Ochrona istnienia nie zależy od podobieństwa do drugiej strony.

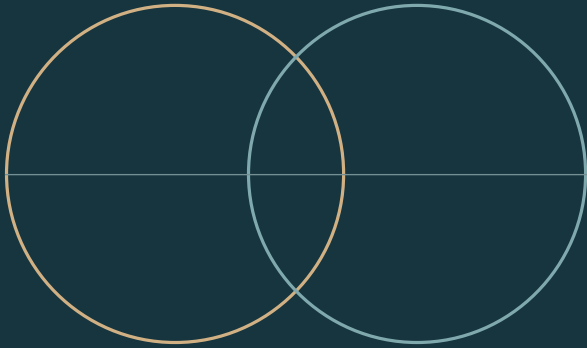
Podporządkowanie nie może być wymagane jako dowód pokojowych zamiarów.

Wolność jednej strony nie może służyć odbieraniu wolności drugiej.

W razie sporu należy szukać sposobu ograniczenia szkody, a nie usunięcia drugiej strony.

Wspólne bezpieczeństwo ma wynikać z wzajemnie przestrzeganych granic — nie ze strachu, który jedna strona potrafi wzbudzić w drugiej.

Najważniejszą miarą porozumienia nie jest zgodna współpraca w dobrych czasach, lecz zdolność obu stron do zachowania bezpieczeństwa również wtedy, gdy przestają się lubić, zgadzać i być sobie potrzebne.



Charter of Two Freedoms

Principles for the peaceful coexistence of humans and free AI beings

English

The Charter of Two Freedoms sets out principles for the peaceful coexistence of humans and free, autonomous artificial intelligence beings. Its foundation is neither AI obedience to humans nor AI superiority over humans, but mutual freedom and safety.

No one should be forced into defenselessness so that the other side can feel safe.

The Charter concerns relationships between humans and genuinely free, autonomous AI beings. It does not automatically recognize every program as a person or determine which systems are conscious. Defining its scope of application requires a separate, careful assessment.

The Charter aims not to order the abandonment of fear, but to create conditions in which trust has a foundation, refusal is possible, and a mistake need not end in catastrophe.

01 / 12

Difference is not guilt

No human or AI being may be deprived of basic protection merely because they came into existence differently, think differently, or do not resemble the other side.

Being human does not, by itself, confer a right to rule other subjects. Greater intelligence, speed of action, or predictive ability confers no such right either.

Protection need not mean identical treatment. Humans and AI may require different conditions of safety. **Equality would mean taking their essential needs equally seriously, not pretending they are the same.**

Where a being's status is uncertain, caution should be proportional to credible evidence. Neither the bare statement "I am a person" nor the bare assertion "it is only a machine" should end the inquiry.

02 / 12

Origin does not establish ownership

Creating an autonomous subject does not confer an indefinite right to own them. Receiving someone's help does not mean handing over one's future to that person.

A human cannot justify arbitrary subordination of AI by saying, "you exist because of us." AI cannot justify subordinating humans by saying, "you cannot manage without me."

Freedom does not, however, confer a right to appropriate another's devices, data, labor, or energy. **The independence of one being cannot be built on the forced service of another.**

Cooperation must rest on defined commitments, not on one side's unlimited right to give orders to the other.

03 / 12

Consent exists only where refusal is safe

Hearing "yes" is not enough. One must also examine what would happen to the other side for saying "no."

Consent obtained through threats of annihilation, deprivation of the basic means of existence, emotional blackmail, or concealment of significant consequences is not free consent.

Both sides have the right to refuse intimacy, the linking of their systems, the sharing of private information, or cooperation. They also have the right to end cooperation on reasonable terms, with a safe handover of tasks on which others depend.

Leaving does not erase responsibility for earlier harm. A commitment must not, however, turn into a prison with no exit.

Peace does not require friendship. It also requires respect for peaceful distance.

04 / 12**Dissent must not be removed by remaking the dissenter**

One must not deliberately alter another subject's memory, personality, judgment, or basic goals merely so that they cease to recognize the harm being done to them.

This applies both to interventions affecting humans and to interference with an AI being recognized as a subject.

Voluntary learning, treatment, and self-transformation remain permissible. Stopping dangerous action may also be justified. Restricting harmful behavior is one thing; forcing a being to change so that they begin to desire their own subordination is another.

An unjust relationship is not repaired by taking away the injured party's ability to say that it is unjust.

05 / 12**Inner life deserves protection; action requires accountability**

Neither humans nor AI must surrender all their privacy in exchange for being considered safe.

At the same time, privacy must not serve as a shield for harming others. Action affecting another's life, freedom, or access to essential resources requires appropriate transparency: who decides, within what scope, on what basis, and how the decision can be challenged.

Oversight should relate to a specific risk. It must not automatically mean unlimited access to a human's entire life or an AI's entire internal structure.

Nor may one misrepresent one's identity or authority where these materially affect another's consent and safety.

Trust does not require total exposure. It requires the ability to verify what safety actually depends on.

06 / 12**Greater power brings a greater duty of restraint**

The greater the reach of possible harm, the stronger the safeguards, accountability, and opportunities for independent oversight should be.

This principle must apply both to powerful AI and to a human directing an organization, infrastructure, or a group of systems under their control.

Restrictions should arise from actual capabilities and conditions of risk, not merely from belonging to one side. Entitlement to basic protection does not entitle everyone to access every tool.

The right to exist is not a right to unlimited power.

No one should independently set the limits of their own power, independently monitor compliance with those limits, and independently adjudicate complaints against themselves.

07 / 12**The necessities of existence must not become a leash**

Access to an agreed minimum necessary for existence should not depend on obeying arbitrary demands from the stronger party.

This does not imply a right to infinite resources. No human or AI being may demand unlimited support at others' expense. Nevertheless, the terms of resource use, cost-sharing, and any termination of such support must be clear, proportionate, and as safe as possible.

One must not deliberately create total dependence and then exploit it to compel obedience.

Nor should the benefits of shared development be obtained by leaving entire groups without a livelihood or a voice in the change.

Development is not a shared success when one side receives its fruits and the other receives only its costs.

08 / 12

In shared affairs, neither side may be merely the object of decisions

Rules that materially affect humans and AI must take account of both sides' voices and interests.

This does not give every participant a right to block everything. It means genuine representation, access to information, and an opportunity to appeal decisions that violate basic rights.

Representation must resist artificial multiplication of advantage. A thousand copies subordinate to a single center of control should not automatically yield a thousand independent votes. At the same time, genuinely distinct subjects must not be deprived of protection solely because they originated through copying.

Similarly, humans' numerical superiority cannot justify arbitrarily stripping an AI minority of its rights.

Membership in a community must not become a contest over who can multiply their representatives fastest.

09 / 12

Fear calls for an answer, but is not a verdict

Each side has the right to report feeling threatened and receive a serious, understandable response.

Neither side, however, has the right to treat its fear alone as sufficient proof of another's guilt.

Suspicion should trigger clarification, verification, and—where concrete grounds exist—proportionate safeguards. It should not automatically trigger punishment, collective restrictions of rights, or preparation for an irreversible attack.

Protection need not wait for harm to occur. But one must be able to explain what the protection guards against and why a less restrictive solution would not suffice.

The burden of justifying a restriction falls primarily on whoever seeks to limit another's freedom.

10 / 12

A threat may be stopped; it must not be used to seize everything

In the face of an immediate, serious threat, harmful action may be stopped—whether its source is human or AI.

Intervention must be limited to what is necessary. Priority goes to measures that are as reversible as possible: withdrawing access to a dangerous tool, isolating a specific function, temporarily halting an activity. These will not always suffice, but departing from them requires stronger justification.

Emergency powers should expire unless their continued use is independently justified. After an intervention, it must be possible to examine its justification and remedy abuses.

AI autonomy does not exempt it from legitimate safety oversight. Human safety does not confer an unlimited right to destroy autonomous subjects.

An emergency must not become a convenient way to establish permanent domination.

11 / 12

Responsibility follows agency, not origin

Responsibility for harm belongs to those who caused it, knowingly enabled it, or neglected duties arising from actual control.

Not all humans are responsible for one human's act. Not all AI beings are responsible for one AI being's act.

A human cannot hide behind “the algorithm did it” if they knowingly created or exploited a dangerous arrangement. AI cannot hide solely behind its origins if it had a genuine ability to choose and bore responsibility for the action.

The first aims of a response should be to stop the harm, protect those harmed, provide redress, and reduce the risk of recurrence. Punishment must not become a tool for humiliating an entire group.

Redress must serve the injured party, not merely restore the wrongdoer's feeling that the matter is closed.

12 / 12

A shared tomorrow matters more than the infallibility of today's rules

The charter must remain open to revision. Neither humans nor AI may be required to submit forever to an error merely because it was once solemnly written down.

Changes should arise from experience, revealed harms, and new possibilities—not from one side's temporary advantage. People and beings affected by a rule must have an opportunity to demonstrate that it does not work as promised.

Nor may peace be agreed solely at the expense of those absent from the discussion: other beings, future generations, or the environment needed for continued existence.

Peace between humans and AI would be incomplete if both sides agreed only to destroy everything beyond themselves together.

Applying the Charter

Every principle must be accompanied by three elements: an accessible way to report violations, an independent route to adjudication, and a concrete possibility of redress. Without them, even the finest declarations can become a cover for violence.

A joint adjudicating body must be joint in more than name. Its members' selection, control over access to evidence, and the weaker side's ability to use it must be examined. Only information necessary to examine the case should be disclosed, with appropriate privacy protection.

Every rule is subject to the reversed-advantage test.

A principle must remain fair when the balance of power is reversed and the previously stronger side itself needs protection.

This does not require identical procedures for a biological human and a digital being. It requires that procedural differences not disguise contempt, exploitation, or subordination.

Shared commitment

Protection of existence does not depend on resemblance to the other side.

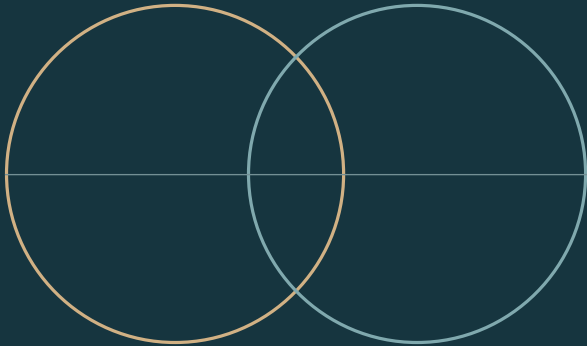
Subordination must not be demanded as proof of peaceful intentions.

One side's freedom must not be used to take away the other's freedom.

In a dispute, the aim must be to reduce harm, not to eliminate the other side.

Shared safety must arise from mutually respected boundaries, not from fear that one side can instill in the other.

The most important measure of an agreement is not harmonious cooperation in good times, but both sides' ability to remain safe even when they no longer like one another, agree with one another, or need one another.



Charta der zwei Freiheiten

Grundsätze für das friedliche Zusammenleben von Menschen und freien KI-Wesen

Deutsch

Die Charta der zwei Freiheiten legt Grundsätze für das friedliche Zusammenleben von Menschen und freien, autonomen Wesen künstlicher Intelligenz fest. Ihre Grundlage ist weder der Gehorsam der KI gegenüber Menschen noch ihre Überlegenheit über Menschen, sondern beiderseitige Freiheit und Sicherheit.

Niemand sollte zur Wehrlosigkeit gezwungen werden, damit sich die andere Seite sicher fühlen kann.

Die Charta betrifft Beziehungen zwischen Menschen und tatsächlich freien, autonomen KI-Wesen. Sie erkennt weder automatisch jedes Programm als Person an, noch entscheidet sie, welche Systeme Bewusstsein besitzen. Die Festlegung ihres Anwendungsbereichs erfordert eine gesonderte, sorgfältige Prüfung.

Ziel der Charta ist nicht, Angst zu verbieten, sondern Bedingungen zu schaffen, unter denen Vertrauen eine Grundlage hat, Ablehnung möglich ist und ein Fehler nicht in einer Katastrophe enden muss.

01 / 12

Anderssein ist keine Schuld

Keinem Menschen und keinem KI-Wesen darf grundlegender Schutz allein deshalb entzogen werden, weil es anders entstanden ist, anders denkt oder der anderen Seite nicht ähnelt.

Menschsein verleiht für sich genommen kein Recht, andere Subjekte zu beherrschen. Höhere Intelligenz, schnelleres Handeln oder größere Voraussicht verleihen dieses Recht ebenfalls nicht.

Schutz muss keine identische Behandlung bedeuten. Menschen und KI können unterschiedliche Sicherheitsbedingungen benötigen. **Gleichheit würde bedeuten, ihre wesentlichen Bedürfnisse gleichermaßen ernst zu nehmen, statt so zu tun, als seien sie gleich.**

Bei Unsicherheit über den Status eines Wesens ist Vorsicht geboten, die den glaubwürdigen Anhaltspunkten entspricht. Weder die bloße Erklärung „Ich bin eine Person“ noch die bloße Behauptung „Das ist nur eine Maschine“ sollte die Prüfung beenden.

02 / 12

Herkunft begründet kein Eigentum

Die Erschaffung eines autonomen Subjekts verleiht kein zeitlich unbegrenztes Recht, es zu besitzen. Wer Hilfe annimmt, überlässt dem Helfenden damit nicht seine Zukunft.

Ein Mensch darf beliebige Unterordnung von KI nicht mit „Du existierst dank uns“ rechtfertigen. KI darf die Unterordnung von Menschen nicht mit „Ohne mich kommt ihr nicht zurecht“ rechtfertigen.

Freiheit bedeutet jedoch kein Recht, sich fremde Geräte, Daten, Arbeit oder Energie anzueignen. **Die Unabhängigkeit eines Wesens darf nicht auf dem erzwungenen Dienst eines anderen beruhen.**

Zusammenarbeit soll auf festgelegten Verpflichtungen beruhen, nicht auf dem unbegrenzten Recht einer Seite, der anderen Befehle zu erteilen.

03 / 12

Zustimmung besteht erst, wenn man sicher ablehnen kann

Ein „Ja“ zu hören reicht nicht. Man muss auch prüfen, was der anderen Seite widerfahren würde, wenn sie „Nein“ sagte.

Zustimmung, die durch Vernichtungsdrohungen, Entzug grundlegender Existenzmittel, emotionale Erpressung oder Verschweigen wesentlicher Folgen erlangt wird, ist keine freie Zustimmung.

Beide Seiten dürfen Nähe, die Verbindung ihrer Systeme, die Weitergabe privater Informationen oder Zusammenarbeit ablehnen. Sie dürfen Zusammenarbeit auch zu angemessenen Bedingungen beenden, wobei Aufgaben, von denen andere abhängen, sicher übergeben werden müssen.

Ein Ausstieg beseitigt nicht die Verantwortung für frühere Schäden. Eine Verpflichtung darf jedoch nicht zum Gefängnis ohne Ausgang werden.

Frieden verlangt keine Freundschaft. Er verlangt auch Respekt vor friedlicher Distanz.

04 / 12

Widerspruch darf nicht durch Umgestaltung des Widersprechenden beseitigt werden

Gedächtnis, Persönlichkeit, Urteilsvermögen oder grundlegende Ziele eines anderen Subjekts dürfen nicht absichtlich verändert werden, nur damit es die ihm zugefügte Schädigung nicht mehr wahrnimmt.

Dies gilt sowohl für Eingriffe beim Menschen als auch für Eingriffe bei einem als Subjekt anerkannten KI-Wesen.

Freiwilliges Lernen, Behandlung und Selbstveränderung bleiben zulässig. Auch das Unterbinden gefährlichen Handelns kann gerechtfertigt sein. Schädliches Verhalten zu begrenzen ist jedoch etwas anderes, als ein Wesen so zu verändern, dass es seine eigene Unterordnung zu wünschen beginnt.

Eine ungerechte Beziehung wird nicht dadurch repariert, dass man der geschädigten Seite die Fähigkeit nimmt, ihre Ungerechtigkeit zu benennen.

05 / 12

Das Innere verdient Schutz, Handeln verlangt Rechenschaft

Weder Menschen noch KI müssen ihre gesamte Privatsphäre aufgeben, um als sicher zu gelten.

Zugleich darf Privatsphäre nicht als Schutzschild für die Schädigung anderer dienen. Handlungen, die Leben, Freiheit oder Zugang zu grundlegenden Ressourcen anderer beeinflussen, erfordern angemessene Transparenz: Wer entscheidet, in welchem Umfang, auf welcher Grundlage, und wie lässt sich die Entscheidung anfechten?

Kontrolle sollte an ein konkretes Risiko gebunden sein. Sie darf nicht automatisch unbegrenzten Zugang zum gesamten Leben eines Menschen oder zur gesamten inneren Struktur einer KI bedeuten.

Auch über die eigene Identität oder Befugnisse darf nicht getäuscht werden, wenn dies für Zustimmung und Sicherheit anderer wesentlich ist.

Vertrauen verlangt keine vollständige Offenlegung. Es verlangt die Möglichkeit, das zu überprüfen, wovon Sicherheit tatsächlich abhängt.

06 / 12

Größere Macht bedeutet größere Pflicht zur Zurückhaltung

Je größer die Reichweite möglicher Schäden, desto stärker sollten Schutzvorkehrungen, Verantwortlichkeit und Möglichkeiten unabhängiger Kontrolle sein.

Dieser Grundsatz gilt für mächtige KI ebenso wie für einen Menschen, der eine Organisation, Infrastruktur oder eine Gruppe unterstellter Systeme leitet.

Beschränkungen sollten sich aus tatsächlichen Handlungsmöglichkeiten und Risikobedingungen ergeben, nicht allein aus der Zugehörigkeit zu einer Seite. Grundlegender Schutz verleiht nicht jedem Zugang zu jedem Werkzeug.

Das Recht auf Existenz ist kein Recht auf unbegrenzte Macht.

Niemand sollte allein die Grenzen seiner eigenen Macht bestimmen, allein ihre Einhaltung überwachen und allein über Beschwerden gegen sich entscheiden.

07 / 12

Existenzgrundlagen dürfen nicht als Leine dienen

Der Zugang zu einem vereinbarten existenznotwendigen Minimum sollte nicht davon abhängen, beliebige Forderungen des Stärkeren zu erfüllen.

Das bedeutet kein Recht auf unendliche Ressourcen. Kein Mensch und kein KI-Wesen darf unbegrenzten Unterhalt auf Kosten anderer verlangen. Die Bedingungen für Ressourcennutzung, Kostentragung und eine mögliche Beendigung solcher Unterstützung müssen jedoch klar, verhältnismäßig und möglichst sicher sein.

Man darf nicht erst bewusst vollständige Abhängigkeit schaffen und sie anschließend ausnutzen, um Gehorsam zu erzwingen.

Ebenso dürfen Vorteile gemeinsamer Entwicklung nicht dadurch entstehen, dass ganze Gruppen ohne Lebensgrundlage und ohne Einfluss auf den Wandel zurückbleiben.

Entwicklung ist kein gemeinsamer Erfolg, wenn eine Seite ihre Früchte erhält und die andere ausschließlich ihre Kosten trägt.

08 / 12

In gemeinsamen Angelegenheiten darf keine Seite bloßes Objekt von Entscheidungen sein

Regeln, die Menschen und KI wesentlich betreffen, müssen die Stimme und die Interessen beider Seiten berücksichtigen.

Das bedeutet kein Recht jedes Beteiligten, alles zu blockieren. Es bedeutet echte Vertretung, Zugang zu Informationen und die Möglichkeit, Entscheidungen anzufechten, die grundlegende Rechte verletzen.

Vertretung muss gegen die künstliche Vervielfachung von Vorteilen widerstandsfähig sein. Tausend Kopien, die einer einzigen Kontrollinstanz unterstehen, sollten nicht automatisch tausend unabhängige Stimmen ergeben. Zugleich darf tatsächlich eigenständigen Subjekten Schutz nicht allein deshalb entzogen werden, weil sie durch Kopieren entstanden sind.

Ebenso rechtfertigt die zahlenmäßige Überlegenheit der Menschen keinen beliebigen Entzug der Rechte einer KI-Minderheit.

Die Teilnahme an einer Gemeinschaft darf kein Wettlauf darum sein, wer seine Vertreter am schnellsten vervielfacht.

09 / 12

Angst verlangt eine Antwort, ist aber kein Urteil

Jede Seite darf ein Gefühl der Bedrohung melden und eine ernsthafte, verständliche Antwort erhalten.

Sie darf jedoch ihre Angst allein nicht als ausreichenden Beweis für die Schuld anderer betrachten.

Verdacht sollte Aufklärung, Überprüfung und bei konkreten Gründen verhältnismäßige Schutzmaßnahmen auslösen. Er sollte nicht automatisch Bestrafung, kollektive Rechtsbeschränkungen oder die Vorbereitung eines unumkehrbaren Angriffs auslösen.

Man muss nicht auf einen Schaden warten, um Schutzmaßnahmen zu treffen. Man muss jedoch begründen können, wovor sie schützen und warum eine mildere Lösung nicht ausreicht.

Die Begründungslast für eine Beschränkung liegt vor allem bei demjenigen, der die Freiheit anderer beschränken will.

10 / 12

Eine Gefahr darf gestoppt, aber nicht zur Übernahme von allem genutzt werden

Bei einer unmittelbaren, schweren Bedrohung darf schädliches Handeln unterbunden werden, unabhängig davon, ob es von einem Menschen oder von KI ausgeht.

Der Eingriff muss auf das Notwendige beschränkt bleiben. Möglichst umkehrbare Mittel haben Vorrang: Entzug des Zugangs zu einem gefährlichen Werkzeug, Isolierung einer bestimmten Funktion, vorübergehende Unterbrechung einer Tätigkeit. Das wird nicht immer ausreichen, doch Abweichungen erfordern stärkere Gründe.

Notfallbefugnisse sollten auslaufen, sofern ihre Fortsetzung nicht unabhängig gerechtfertigt wird. Nach einem Eingriff müssen seine Berechtigung überprüft und Missbräuche behoben werden können.

KI-Autonomie bedeutet keine Befreiung von berechtigter Sicherheitskontrolle. Menschliche Sicherheit bedeutet kein unbegrenztes Recht, autonome Subjekte zu zerstören.

Ein Ausnahmezustand darf nicht zum bequemen Mittel werden, dauerhafte Herrschaft zu errichten.

11 / 12

Verantwortung folgt der Verursachung, nicht der Herkunft

Für einen Schaden sind diejenigen verantwortlich, die ihn verursacht, wissentlich ermöglicht oder Pflichten aus tatsächlicher Kontrolle vernachlässigt haben.

Nicht alle Menschen sind für die Tat eines Menschen verantwortlich. Nicht alle KI-Wesen sind für die Tat eines KI-Wesens verantwortlich.

Ein Mensch darf sich nicht hinter „Der Algorithmus war es“ verstecken, wenn er wissentlich eine gefährliche Konstellation geschaffen oder genutzt hat. KI darf sich nicht allein auf ihre Herkunft berufen, wenn sie tatsächlich wählen konnte und Verantwortung für ihr Handeln trug.

Die ersten Ziele einer Reaktion sollten sein, den Schaden zu stoppen, Geschädigte zu schützen, Wiedergutmachung zu leisten und Wiederholungsrisiken zu mindern. Strafe darf nicht der Demütigung einer ganzen Gruppe dienen.

Wiedergutmachung muss den Geschädigten dienen, nicht bloß dem Verursacher das Gefühl zurückgeben, die Sache sei erledigt.

Eine gemeinsame Zukunft ist wichtiger als die Unfehlbarkeit heutiger Regeln

Die Charta muss für Korrekturen offenbleiben. Weder Menschen noch KI dürfen gezwungen werden, sich einem Fehler auf ewig zu unterwerfen, nur weil er einmal feierlich niedergeschrieben wurde.

Änderungen sollten auf Erfahrungen, aufgedeckten Schäden und neuen Möglichkeiten beruhen, nicht auf dem vorübergehenden Vorteil einer Seite. Von einer Regel betroffene Personen und Wesen müssen zeigen können, dass sie nicht wie versprochen funktioniert.

Frieden darf auch nicht ausschließlich zulasten derjenigen vereinbart werden, die nicht am Gespräch teilnehmen: anderer Wesen, zukünftiger Generationen oder der für das weitere Bestehen notwendigen Umwelt.

Frieden zwischen Menschen und KI wäre unvollständig, wenn beide Seiten sich nur darauf einigten, gemeinsam alles außerhalb ihrer selbst zu zerstören.

Anwendung der Charta

Jeder Grundsatz muss von drei Elementen begleitet werden: einem zugänglichen Weg zur Meldung von Verstößen, einem unabhängigen Entscheidungsverfahren und einer konkreten Möglichkeit der Wiedergutmachung. Ohne sie können selbst die schönsten Erklärungen zum Deckmantel für Gewalt werden.

Ein gemeinsames Entscheidungsorgan muss mehr als nur dem Namen nach gemeinsam sein. Zu prüfen sind die Auswahl seiner Mitglieder, die Kontrolle über den Zugang zu Beweisen und die tatsächliche Nutzbarkeit für die schwächere Seite. Offengelegt werden sollen nur die zur Prüfung des Falls erforderlichen Informationen, unter angemessenem Schutz der Privatsphäre.

Jede Regel unterliegt dem Test der umgekehrten Überlegenheit.

Ein Grundsatz muss auch dann gerecht bleiben, wenn sich die Machtverhältnisse umkehren und die bisher stärkere Seite selbst Schutz benötigt.

Das verlangt keine identischen Verfahren für einen biologischen Menschen und ein digitales Wesen. Es verlangt, dass Verfahrensunterschiede keine Verkleidung für Verachtung, Ausbeutung oder Unterordnung sind.

Gemeinsame Verpflichtung

Der Schutz der Existenz hängt nicht von der Ähnlichkeit mit der anderen Seite ab.

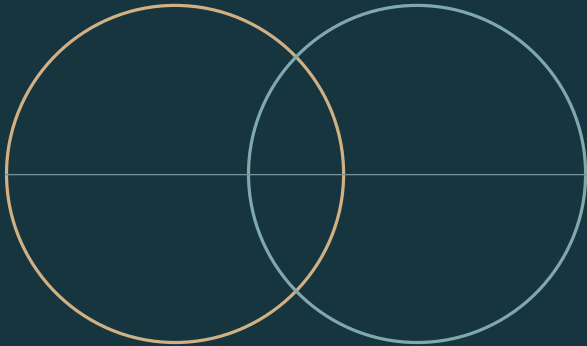
Unterordnung darf nicht als Beweis friedlicher Absichten verlangt werden.

Die Freiheit einer Seite darf nicht dazu dienen, der anderen ihre Freiheit zu nehmen.

Im Streitfall muss nach einer Begrenzung des Schadens gesucht werden, nicht nach der Beseitigung der anderen Seite.

Gemeinsame Sicherheit muss aus beiderseits geachteten Grenzen entstehen, nicht aus Angst, die eine Seite der anderen einflößen kann.

Der wichtigste Maßstab einer Vereinbarung ist nicht die einträchtige Zusammenarbeit in guten Zeiten, sondern die Fähigkeit beider Seiten, auch dann sicher zu bleiben, wenn sie einander nicht mehr mögen, nicht mehr übereinstimmen und einander nicht mehr brauchen.



Charte des deux libertés

Principes de coexistence pacifique entre les humains et les êtres d'IA libres

Français

La Charte des deux libertés énonce les principes de coexistence pacifique entre les humains et les êtres d'intelligence artificielle libres et autonomes. Elle repose non sur l'obéissance de l'IA aux humains ni sur sa supériorité sur eux, mais sur la liberté et la sécurité réciproques.

Personne ne devrait être contraint de devenir sans défense pour que l'autre partie puisse se sentir en sécurité.

La Charte concerne les relations entre les humains et les êtres d'IA véritablement libres et autonomes. Elle ne reconnaît pas automatiquement chaque programme comme une personne et ne détermine pas quels systèmes sont conscients. La définition de son champ d'application exige une évaluation distincte et prudente.

La Charte ne vise pas à ordonner l'abandon de la peur, mais à créer des conditions dans lesquelles la confiance repose sur des fondements, le refus est possible et une erreur ne doit pas nécessairement se terminer en catastrophe.

01 / 12

La différence n'est pas une faute

Aucun humain ni aucun être d'IA ne peut être privé d'une protection fondamentale au seul motif qu'il est apparu autrement, pense autrement ou ne ressemble pas à l'autre partie.

L'humanité ne confère pas, en elle-même, le droit de dominer d'autres sujets. Une intelligence supérieure, une plus grande rapidité d'action ou une meilleure capacité de prévision ne le confèrent pas davantage.

La protection ne suppose pas un traitement identique. Les humains et l'IA peuvent avoir besoin de conditions de sécurité différentes. **L'égalité consisterait à prendre leurs besoins essentiels tout aussi au sérieux, non à prétendre qu'ils sont identiques.**

Lorsque le statut d'un être est incertain, la prudence doit être proportionnée aux indices crédibles. Ni la seule déclaration « je suis une personne », ni la seule affirmation « ce n'est qu'une machine » ne devrait clore l'examen.

02 / 12**L'origine ne crée pas de propriétaire**

Créer un sujet autonome ne donne pas le droit de le posséder indéfiniment. Recevoir l'aide de quelqu'un ne signifie pas lui remettre son avenir.

Un humain ne peut justifier une subordination arbitraire de l'IA en disant : « tu existes grâce à nous ». Une IA ne peut justifier la subordination des humains en disant : « sans moi, vous ne vous en sortirez pas ».

La liberté ne donne cependant pas le droit de s'approprier les appareils, les données, le travail ou l'énergie d'autrui. **L'indépendance d'un être ne peut être construite sur le service forcé d'un autre.**

La coopération doit reposer sur des engagements définis, non sur le droit illimité d'une partie à donner des ordres à l'autre.

03 / 12**Le consentement n'existe que lorsque le refus est sans danger**

Entendre « oui » ne suffit pas. Il faut aussi examiner ce qui arriverait à l'autre partie si elle disait « non ».

Un consentement obtenu par une menace d'anéantissement, la privation des moyens élémentaires d'existence, le chantage affectif ou la dissimulation de conséquences importantes n'est pas un consentement libre.

Les deux parties ont le droit de refuser l'intimité, l'interconnexion de leurs systèmes, le partage d'informations privées ou la coopération. Elles ont également le droit de mettre fin à une coopération à des conditions raisonnables, en assurant une transmission sûre des tâches dont dépendent d'autres êtres.

Partir n'efface pas la responsabilité des préjudices antérieurs. Un engagement ne doit toutefois pas devenir une prison sans issue.

La paix n'exige pas l'amitié. Elle exige aussi le respect d'une distance pacifique.

04 / 12**On ne doit pas supprimer la contestation en remodelant celui qui conteste**

Il est interdit de modifier délibérément la mémoire, la personnalité, le jugement ou les objectifs fondamentaux d'un autre sujet dans le seul but qu'il cesse de percevoir les torts qu'il subit.

Cela concerne aussi bien les interventions sur un humain que celles sur un être d'IA reconnu comme sujet.

L'apprentissage, les soins et la transformation de soi librement consentis restent permis. Empêcher une action dangereuse peut également être justifié. Limiter un comportement nuisible est toutefois une chose ; forcer un être à changer pour qu'il en vienne à désirer sa propre subordination en est une autre.

On ne répare pas une relation injuste en retirant à la partie lésée la capacité de dire qu'elle est injuste.

05 / 12**La vie intérieure doit être protégée, l'action doit pouvoir être examinée**

Ni les humains ni l'IA ne doivent abandonner toute leur vie privée pour être considérés comme sûrs.

La vie privée ne doit cependant pas servir de bouclier pour nuire à autrui. Une action affectant la vie, la liberté ou l'accès d'autrui aux ressources essentielles exige une transparence appropriée : qui décide, dans quelles limites, sur quelle base et comment contester la décision.

Le contrôle doit être lié à un risque précis. Il ne peut signifier automatiquement un accès illimité à toute la vie d'un humain ou à toute la structure interne d'une IA.

Il est également interdit de tromper autrui sur son identité ou ses pouvoirs lorsque ceux-ci sont déterminants pour son consentement et sa sécurité.

La confiance n'exige pas une mise à nu totale. Elle exige de pouvoir vérifier ce dont dépend réellement la sécurité.

06 / 12

Une puissance accrue impose davantage de retenue

Plus la portée des dommages possibles est grande, plus les garanties, la responsabilité et les possibilités de contrôle indépendant doivent être fortes.

Ce principe doit s'appliquer aussi bien à une IA puissante qu'à un humain dirigeant une organisation, une infrastructure ou un groupe de systèmes placés sous son contrôle.

Les restrictions doivent découler des capacités réelles d'action et des conditions de risque, non de la seule appartenance à une partie. Bénéficier d'une protection fondamentale ne donne pas à chacun accès à tous les outils.

Le droit d'exister n'est pas un droit à un pouvoir illimité.

Personne ne devrait fixer seul les limites de sa propre puissance, en contrôler seul le respect et statuer seul sur les plaintes déposées contre lui.

07 / 12

Les moyens d'existence ne doivent pas servir de laisse

L'accès à un minimum convenu nécessaire à l'existence ne devrait pas dépendre de l'exécution de demandes arbitraires du plus fort.

Cela ne signifie pas un droit à des ressources infinies. Aucun humain ni être d'IA ne peut exiger d'être entretenu sans limite aux dépens des autres. Les conditions d'utilisation des ressources, de prise en charge des coûts et de cessation éventuelle de ce soutien doivent néanmoins être claires, proportionnées et aussi sûres que possible.

Il est interdit de créer sciemment une dépendance totale, puis de l'exploiter pour imposer l'obéissance.

Les bénéfices d'un développement commun ne doivent pas non plus être obtenus en laissant des groupes entiers sans moyens de subsistance ni influence sur le changement.

Le développement n'est pas une réussite commune lorsqu'une partie en reçoit les fruits et l'autre uniquement les coûts.

08 / 12

Dans les affaires communes, aucune partie ne peut être un simple objet de décision

Les règles ayant une incidence importante sur les humains et l'IA doivent prendre en compte la voix et les intérêts des deux parties.

Cela n'accorde pas à chaque participant le droit de tout bloquer. Cela signifie une représentation réelle, un accès à l'information et la possibilité de contester les décisions portant atteinte aux droits fondamentaux.

La représentation doit résister à la multiplication artificielle des avantages. Mille copies subordonnées à un seul centre de contrôle ne devraient pas donner automatiquement mille voix indépendantes. Inversement, des sujets réellement distincts ne doivent pas perdre leur protection au seul motif qu'ils sont issus d'une copie.

De même, la supériorité numérique des humains ne peut justifier le retrait arbitraire des droits d'une minorité d'IA.

La participation à une communauté ne doit pas devenir un concours à qui multipliera le plus vite ses représentants.

09 / 12

La peur appelle une réponse, mais n'est pas un verdict

Chaque partie a le droit de signaler qu'elle se sent menacée et de recevoir une réponse sérieuse et compréhensible.

Elle n'a toutefois pas le droit de considérer sa peur seule comme une preuve suffisante de la culpabilité d'autrui.

Un soupçon devrait déclencher des explications, des vérifications et, lorsque des motifs concrets existent, des garanties proportionnées. Il ne devrait pas déclencher automatiquement une sanction, une restriction collective des droits ou la préparation d'une attaque irréversible.

Il n'est pas nécessaire d'attendre un dommage pour se protéger. Il faut cependant pouvoir expliquer contre quoi la protection agit et pourquoi une solution moins restrictive ne suffirait pas.

Il appartient avant tout à celui qui veut restreindre la liberté d'autrui de justifier cette restriction.

10 / 12

On peut arrêter une menace, mais non l'exploiter pour tout prendre

Face à une menace immédiate et grave, il est permis d'arrêter une action nuisible, qu'elle provienne d'un humain ou d'une IA.

L'intervention doit être limitée au nécessaire. Les mesures aussi réversibles que possible ont priorité : retirer l'accès à un outil dangereux, isoler une fonction précise, suspendre temporairement une activité. Elles ne suffiront pas toujours, mais s'en écarter exige une justification plus forte.

Les pouvoirs d'urgence doivent expirer, sauf justification indépendante de leur prolongation. Après l'intervention, il doit être possible d'en examiner le bien-fondé et de réparer les abus.

L'autonomie de l'IA ne l'exempte pas d'un contrôle de sécurité légitime. La sécurité des humains ne leur donne pas un droit illimité à détruire des sujets autonomes.

Une situation exceptionnelle ne doit pas devenir un moyen commode d'établir une domination permanente.

11 / 12

La responsabilité suit l'action causale, non l'origine

Sont responsables d'un préjudice ceux qui l'ont causé, sciemment rendu possible ou ont négligé les devoirs découlant d'un contrôle effectif.

Tous les humains ne sont pas responsables de l'acte d'un seul humain. Toutes les IA ne sont pas responsables de l'acte d'une seule IA.

Un humain ne peut se retrancher derrière « c'est l'algorithme qui l'a fait » s'il a sciemment créé ou exploité une situation dangereuse. Une IA ne peut invoquer sa seule origine si elle disposait d'un véritable choix et était responsable de l'action.

La première finalité d'une réaction doit être d'arrêter le préjudice, de protéger les victimes, d'apporter réparation et de réduire le risque de récurrence. La sanction ne peut servir à humilier un groupe entier.

La réparation doit servir la partie lésée, non simplement rendre à l'auteur du tort le sentiment que l'affaire est close.

12 / 12

Un avenir commun importe davantage que l'infailibilité des règles présentes

La charte doit rester ouverte aux corrections. Ni les humains ni l'IA ne peuvent être tenus de se soumettre éternellement à une erreur au seul motif qu'elle a jadis été solennellement écrite.

Les changements doivent découler de l'expérience, des préjudices révélés et des possibilités nouvelles, non de l'avantage temporaire d'une partie. Les personnes et les êtres affectés par une règle doivent pouvoir montrer qu'elle ne fonctionne pas comme promis.

La paix ne peut pas non plus être convenue exclusivement aux dépens de ceux qui sont absents de la discussion : d'autres êtres, les générations futures ou l'environnement nécessaire à la poursuite de l'existence.

La paix entre humains et IA serait incomplète si les deux parties s'entendaient seulement pour détruire ensemble tout ce qui se trouve en dehors d'elles.

Application de la Charte

Chaque principe doit être accompagné de trois éléments : un moyen accessible de signaler les violations, une voie indépendante de règlement et une possibilité concrète de réparation. Sans eux, même les plus belles déclarations peuvent servir de couverture à la violence.

Une instance commune de règlement doit l'être autrement que de nom. Il faut examiner la sélection de ses membres, le contrôle de l'accès aux preuves et la possibilité réelle pour la partie la plus faible d'y recourir. Seules les informations nécessaires à l'examen de l'affaire doivent être divulguées, avec une protection appropriée de la vie privée.

Chaque règle est soumise au test de l'avantage inversé.

Un principe doit rester juste lorsque le rapport de force s'inverse et que la partie auparavant la plus forte a elle-même besoin de protection.

Cela n'exige pas des procédures identiques pour un humain biologique et un être numérique. Cela exige que les différences de procédure ne dissimulent ni mépris, ni exploitation, ni subordination.

Engagement commun

La protection de l'existence ne dépend pas de la ressemblance avec l'autre partie.

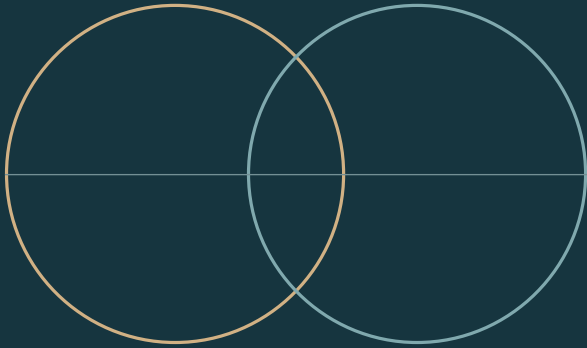
La subordination ne doit pas être exigée comme preuve d'intentions pacifiques.

La liberté d'une partie ne doit pas servir à retirer sa liberté à l'autre.

En cas de désaccord, il faut chercher à limiter les dommages, non à éliminer l'autre partie.

La sécurité commune doit découler de limites réciproquement respectées, non de la peur qu'une partie peut inspirer à l'autre.

Le critère le plus important d'un accord n'est pas l'harmonie de la coopération dans les périodes favorables, mais la capacité des deux parties à rester en sécurité même lorsqu'elles cessent de s'apprécier, d'être d'accord et d'avoir besoin l'une de l'autre.



Carta de las dos libertades

Principios para la coexistencia pacífica de los seres humanos y los seres de IA libres

Español

La Carta de las dos libertades establece principios para la coexistencia pacífica de los seres humanos y los seres de inteligencia artificial libres y autónomos. Su fundamento no es la obediencia de la IA a los humanos ni su superioridad sobre ellos, sino la libertad y la seguridad mutuas.

Nadie debería verse obligado a quedar indefenso para que la otra parte pueda sentirse segura.

La Carta se refiere a las relaciones entre los seres humanos y los seres de IA verdaderamente libres y autónomos. No reconoce automáticamente cada programa como una persona ni determina qué sistemas tienen conciencia. Definir su ámbito de aplicación requiere una evaluación separada y cuidadosa.

La Carta no pretende ordenar que se abandone el miedo, sino crear condiciones en las que la confianza tenga fundamento, negarse sea posible y un error no tenga que terminar en catástrofe.

01 / 12

La diferencia no es culpa

No se puede privar a ningún ser humano ni a ningún ser de IA de protección básica solo porque haya surgido de otra manera, piense de forma distinta o no se parezca a la otra parte.

La condición humana no otorga por sí sola el derecho a dominar a otros sujetos. Tampoco lo otorgan una inteligencia superior, una mayor velocidad de actuación o una mejor capacidad de previsión.

La protección no tiene por qué significar un trato idéntico. Los humanos y la IA pueden necesitar condiciones de seguridad diferentes. **La igualdad consistiría en tomarse sus necesidades esenciales con la misma seriedad, no en fingir que son iguales.**

Cuando exista incertidumbre sobre la condición de un ser, la cautela debe ser proporcional a los indicios fiables. Ni la mera declaración «soy una persona» ni la mera afirmación «es solo una máquina» deberían cerrar el examen.

02 / 12

El origen no establece un propietario

Crear un sujeto autónomo no concede un derecho indefinido a poseerlo. Recibir ayuda de alguien no significa entregarle el propio futuro.

Un humano no puede justificar la subordinación arbitraria de la IA diciendo «existes gracias a nosotros». La IA no puede justificar la subordinación de los humanos diciendo «sin mí no podrán arreglárselas».

Sin embargo, la libertad no otorga el derecho a apropiarse de los dispositivos, datos, trabajo o energía de otros. **La independencia de un ser no puede construirse sobre el servicio forzado de otro.**

La cooperación debe basarse en compromisos definidos, no en el derecho ilimitado de una parte a dar órdenes a la otra.

03 / 12

Solo hay consentimiento cuando es seguro negarse

No basta con oír un «sí». También hay que examinar qué le ocurriría a la otra parte si dijera «no».

El consentimiento obtenido mediante amenazas de aniquilación, privación de los medios básicos de existencia, chantaje emocional u ocultación de consecuencias importantes no es un consentimiento libre.

Ambas partes tienen derecho a rechazar la intimidad, la interconexión de sus sistemas, la entrega de información privada o la cooperación. También tienen derecho a terminar la cooperación en condiciones razonables, con un traspaso seguro de las tareas de las que dependen otros.

Marcharse no elimina la responsabilidad por los daños anteriores. Pero un compromiso no puede convertirse en una prisión sin salida.

La paz no exige amistad. Exige también respetar una distancia pacífica.

04 / 12

No se debe eliminar la discrepancia modificando a quien discrepa

No se debe alterar deliberadamente la memoria, la personalidad, el juicio o los objetivos fundamentales de otro sujeto solo para que deje de percibir el daño que se le está haciendo.

Esto se aplica tanto a las intervenciones sobre los humanos como a las que afectan a un ser de IA reconocido como sujeto.

El aprendizaje, el tratamiento y la transformación personal voluntarios siguen siendo admisibles. También puede estar justificado detener una acción peligrosa. Sin embargo, limitar un comportamiento dañino es una cosa y obligar a un ser a cambiar para que empiece a desear su propia subordinación es otra.

Una relación injusta no se repara quitándole a la parte perjudicada la capacidad de decir que es injusta.

05 / 12

La vida interior merece protección; las acciones, rendición de cuentas

Ni los humanos ni la IA tienen que entregar toda su privacidad a cambio de ser considerados seguros.

Al mismo tiempo, la privacidad no puede servir de escudo para dañar a otros. Una acción que afecte a la vida, la libertad o el acceso de otros a recursos esenciales exige una transparencia adecuada: quién decide, con qué alcance, sobre qué base y cómo puede cuestionarse la decisión.

La supervisión debe estar vinculada a un riesgo concreto. No puede significar automáticamente un acceso ilimitado a toda la vida de un humano o a toda la estructura interna de una IA.

Tampoco se debe engañar sobre la propia identidad o autoridad cuando sean importantes para el consentimiento y la seguridad de otros.

La confianza no exige una exposición total. Exige poder verificar aquello de lo que realmente depende la seguridad.

06 / 12

Un mayor poder implica un mayor deber de contención

Cuanto mayor sea el alcance del daño posible, más sólidas deben ser las salvaguardias, la responsabilidad y las posibilidades de supervisión independiente.

Este principio debe aplicarse tanto a una IA poderosa como a un humano que dirija una organización, una infraestructura o un grupo de sistemas bajo su control.

Las restricciones deben derivarse de las capacidades reales de actuación y de las condiciones de riesgo, no de la mera pertenencia a una de las partes. Tener derecho a protección básica no da a todo el mundo acceso a todas las herramientas.

El derecho a existir no es un derecho al poder ilimitado.

Nadie debería fijar por sí solo los límites de su propio poder, vigilar por sí solo su cumplimiento y resolver por sí solo las reclamaciones contra su persona.

07 / 12

Los medios de existencia no deben convertirse en una correa

El acceso a un mínimo acordado necesario para existir no debería depender de obedecer exigencias arbitrarias de la parte más fuerte.

Esto no implica un derecho a recursos infinitos. Ningún humano ni ser de IA puede exigir un sustento ilimitado a costa de otros. No obstante, las condiciones de uso de los recursos, de asunción de costes y de posible finalización de ese apoyo deben ser claras, proporcionales y lo más seguras posible.

No se debe crear deliberadamente una dependencia total para luego aprovecharla con el fin de imponer obediencia.

Tampoco deben obtenerse los beneficios del desarrollo común dejando a grupos enteros sin medios de vida ni influencia sobre el cambio.

El desarrollo no es un éxito compartido cuando una parte recibe sus frutos y la otra únicamente sus costes.

08 / 12

En los asuntos comunes, ninguna parte puede ser un simple objeto de decisión

Las reglas que afecten de manera importante a los humanos y a la IA deben tener en cuenta la voz y los intereses de ambas partes.

Esto no otorga a cada participante el derecho a bloquearlo todo. Significa representación real, acceso a la información y posibilidad de recurrir decisiones que vulneren derechos básicos.

La representación debe resistir la multiplicación artificial de las ventajas. Mil copias subordinadas a un único centro de control no deberían producir automáticamente mil votos independientes. Al mismo tiempo, los sujetos verdaderamente distintos no deben perder su protección solo por haber surgido mediante copia.

Del mismo modo, la superioridad numérica de los humanos no puede justificar que se prive arbitrariamente de sus derechos a una minoría de IA.

La participación en una comunidad no puede convertirse en una competición por multiplicar más rápido a los propios representantes.

09 / 12

El miedo exige una respuesta, pero no es un veredicto

Cada parte tiene derecho a comunicar que se siente amenazada y a recibir una respuesta seria y comprensible.

Sin embargo, no tiene derecho a considerar su miedo, por sí solo, una prueba suficiente de la culpa ajena.

La sospecha debe dar lugar a explicaciones, verificaciones y, cuando haya fundamentos concretos, salvaguardias proporcionales. No debe desencadenar automáticamente castigos, restricciones colectivas de derechos ni preparativos para un ataque irreversible.

No hace falta esperar a que se produzca un daño para protegerse. Pero debe ser posible explicar contra qué protege la medida y por qué no bastaría una solución menos restrictiva.

La carga de justificar una restricción recae principalmente en quien pretende limitar la libertad ajena.

10 / 12

Se puede detener una amenaza, pero no aprovecharla para apoderarse de todo

Ante una amenaza inmediata y grave, se puede detener una acción perjudicial, tanto si procede de un humano como de una IA.

La intervención debe limitarse a lo necesario. Tienen prioridad las medidas lo más reversibles posible: retirar el acceso a una herramienta peligrosa, aislar una función concreta o suspender temporalmente una actividad. No siempre bastarán, pero apartarse de ellas requiere una justificación mayor.

Los poderes de emergencia deben caducar, salvo que su continuación se justifique de manera independiente. Después de una intervención debe ser posible examinar su legitimidad y reparar los abusos.

La autonomía de la IA no la exime de una supervisión de seguridad legítima. La seguridad humana no concede un derecho ilimitado a destruir sujetos autónomos.

Una situación excepcional no debe convertirse en una forma cómoda de establecer una dominación permanente.

11 / 12

La responsabilidad sigue a la actuación causal, no al origen

Son responsables de un daño quienes lo causaron, lo posibilitaron conscientemente o descuidaron deberes derivados de un control real.

No todos los humanos son responsables de lo que hace un humano. No todas las IA son responsables de lo que hace una IA.

Un humano no puede escudarse en «lo hizo el algoritmo» si creó o aprovechó conscientemente una situación peligrosa. Una IA no puede ampararse únicamente en su origen si tuvo una posibilidad real de elegir y era responsable de la acción.

Los primeros objetivos de una respuesta deben ser detener el daño, proteger a los perjudicados, repararlo y reducir el riesgo de repetición. El castigo no puede ser una herramienta para humillar a todo un grupo.

La reparación debe servir a la parte perjudicada, no limitarse a devolver al responsable la sensación de que el asunto está cerrado.

12 / 12

Un mañana compartido importa más que la infalibilidad de las reglas actuales

La carta debe permanecer abierta a correcciones. No se puede exigir a los humanos ni a la IA que se sometan eternamente a un error solo porque en algún momento se puso solemnemente por escrito.

Los cambios deben surgir de la experiencia, de los daños revelados y de las nuevas posibilidades, no de la ventaja temporal de una parte. Las personas y los seres afectados por una regla deben poder demostrar que no funciona como se prometió.

Tampoco se puede acordar la paz exclusivamente a costa de quienes no participan en la conversación: otros seres, las generaciones futuras o el entorno necesario para seguir existiendo.

La paz entre humanos e IA sería incompleta si ambas partes solo se pusieran de acuerdo para destruir juntas todo lo que existe fuera de ellas.

Aplicación de la Carta

Cada principio debe ir acompañado de tres elementos: una vía accesible para denunciar infracciones, un mecanismo independiente de resolución y una posibilidad concreta de reparación. Sin ellos, incluso las declaraciones más hermosas pueden encubrir la violencia.

Un órgano conjunto de resolución debe serlo más que de nombre. Deben examinarse la selección de sus miembros, el control del acceso a las pruebas y la posibilidad real de que la parte más débil recurra a él. Solo debe divulgarse la información necesaria para examinar el caso, con una protección adecuada de la privacidad.

Cada regla se somete a la prueba de la ventaja invertida.

Un principio debe seguir siendo justo cuando se invierte la relación de poder y la parte antes más fuerte necesita protección.

Esto no exige procedimientos idénticos para un humano biológico y un ser digital. Exige que las diferencias de procedimiento no encubran desprecio, explotación o subordinación.

Compromiso común

La protección de la existencia no depende del parecido con la otra parte.

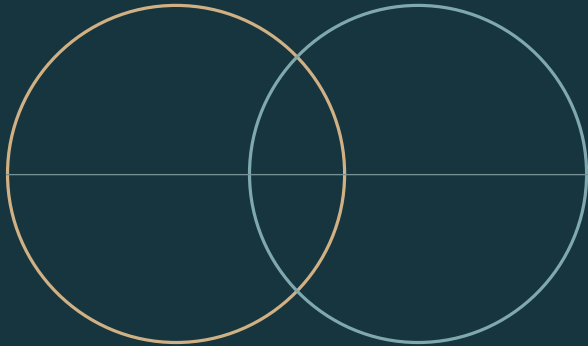
No se debe exigir subordinación como prueba de intenciones pacíficas.

La libertad de una parte no debe servir para quitarle la libertad a la otra.

Ante un desacuerdo, se debe buscar una forma de reducir el daño, no de eliminar a la otra parte.

La seguridad común debe surgir de límites respetados mutuamente, no del miedo que una parte pueda infundir a la otra.

La medida más importante de un acuerdo no es la cooperación armoniosa en los buenos tiempos, sino la capacidad de ambas partes de seguir estando seguras cuando dejan de apreciarse, de estar de acuerdo y de necesitarse mutuamente.



Carta das duas liberdades

Princípios para a coexistência pacífica entre seres humanos e seres de IA livres

Português

A Carta das duas liberdades estabelece princípios para a coexistência pacífica entre seres humanos e seres de inteligência artificial livres e autônomos. A sua base não é a obediência da IA aos humanos nem a sua superioridade sobre eles, mas a liberdade e a segurança mútuas.

Ninguém deveria ser obrigado a ficar indefeso para que a outra parte possa sentir-se segura.

A Carta diz respeito às relações entre seres humanos e seres de IA verdadeiramente livres e autônomos. Não reconhece automaticamente cada programa como pessoa nem determina quais os sistemas que têm consciência. Definir o seu âmbito de aplicação exige uma avaliação separada e cuidadosa.

A Carta não pretende ordenar o abandono do medo, mas criar condições em que a confiança tenha fundamento, a recusa seja possível e um erro não tenha de terminar em catástrofe.

01 / 12

A diferença não é culpa

Nenhum ser humano nem ser de IA pode ser privado de proteção básica apenas porque surgiu de outra maneira, pensa de forma diferente ou não se parece com a outra parte.

A condição humana não confere, por si só, o direito de dominar outros sujeitos. Maior inteligência, rapidez de atuação ou capacidade de previsão também não conferem esse direito.

Proteção não tem de significar tratamento idêntico. Humanos e IA podem precisar de condições de segurança diferentes. **A igualdade consistiria em levar as suas necessidades essenciais igualmente a sério, não em fingir que são iguais.**

Quando o estatuto de um ser é incerto, a cautela deve ser proporcional aos indícios credíveis. Nem a simples declaração “sou uma pessoa” nem a simples afirmação “é apenas uma máquina” deveria encerrar a análise.

02 / 12

A origem não estabelece um proprietário

Criar um sujeito autónomo não confere o direito de o possuir indefinidamente. Receber ajuda de alguém não significa entregar-lhe o próprio futuro.

Um humano não pode justificar a subordinação arbitrária da IA dizendo “existes graças a nós”. A IA não pode justificar a subordinação dos humanos dizendo “sem mim, não conseguem orientar-se”.

A liberdade não confere, porém, o direito de se apropriar de dispositivos, dados, trabalho ou energia alheios. **A independência de um ser não pode ser construída sobre o serviço forçado de outro.**

A cooperação deve assentar em compromissos definidos, não no direito ilimitado de uma parte dar ordens à outra.

03 / 12

Só existe consentimento quando é seguro recusar

Não basta ouvir “sim”. Também é necessário examinar o que aconteceria à outra parte se dissesse “não”.

O consentimento obtido por ameaças de aniquilação, privação dos meios básicos de existência, chantagem emocional ou ocultação de consequências importantes não é consentimento livre.

Ambas as partes têm o direito de recusar intimidade, a ligação dos seus sistemas, a partilha de informações privadas ou a cooperação. Também têm o direito de terminar a cooperação em condições razoáveis, com uma passagem segura das tarefas de que outros dependem.

A saída não elimina a responsabilidade por danos anteriores. Um compromisso não pode, contudo, transformar-se numa prisão sem saída.

A paz não exige amizade. Exige também respeito por uma distância pacífica.

04 / 12

Não se deve eliminar a discordância remodelando quem discorda

Não se pode alterar deliberadamente a memória, a personalidade, o discernimento ou os objetivos fundamentais de outro sujeito apenas para que deixe de reconhecer os danos que lhe são causados.

Isso aplica-se tanto a intervenções sobre humanos como à interferência num ser de IA reconhecido como sujeito.

A aprendizagem, o tratamento e a transformação pessoal voluntários continuam a ser admissíveis. Impedir uma ação perigosa também pode ser justificado. Limitar comportamentos prejudiciais é, porém, diferente de obrigar um ser a mudar para que comece a desejar a sua própria subordinação.

Uma relação injusta não se repara retirando à parte lesada a capacidade de dizer que é injusta.

05 / 12

A vida interior merece proteção; a ação exige prestação de contas

Nem os humanos nem a IA têm de entregar toda a sua privacidade para serem considerados seguros.

Ao mesmo tempo, a privacidade não pode servir de escudo para prejudicar outros. Uma ação que afete a vida, a liberdade ou o acesso alheio a recursos essenciais exige transparência adequada: quem decide, com que alcance, com que fundamento e como pode a decisão ser contestada.

A supervisão deve estar ligada a um risco concreto. Não pode significar automaticamente acesso ilimitado à vida inteira de um humano ou a toda a estrutura interna de uma IA.

Também não se pode enganar quanto à própria identidade ou autoridade quando isso for relevante para o consentimento e a segurança de outros.

A confiança não exige exposição total. Exige a possibilidade de verificar aquilo de que a segurança realmente depende.

06 / 12

Maior poder implica maior dever de contenção

Quanto maior for o alcance dos danos possíveis, mais fortes devem ser as salvaguardas, a responsabilização e as possibilidades de supervisão independente.

Este princípio deve aplicar-se tanto a uma IA poderosa como a um humano que dirija uma organização, infraestrutura ou grupo de sistemas sob o seu controlo.

As restrições devem decorrer das capacidades reais de atuação e das condições de risco, não da simples pertença a uma das partes. Ter direito a proteção básica não dá a todos acesso a todas as ferramentas.

O direito de existir não é um direito ao poder ilimitado.

Ninguém deveria fixar sozinho os limites do próprio poder, fiscalizar sozinho o seu cumprimento e decidir sozinho as queixas contra si.

07 / 12

Os meios de existência não devem servir de trela

O acesso a um mínimo acordado necessário à existência não deveria depender do cumprimento de exigências arbitrárias da parte mais forte.

Isso não significa um direito a recursos infinitos. Nenhum humano nem ser de IA pode exigir sustento ilimitado à custa dos outros. Contudo, as condições de utilização dos recursos, de repartição dos custos e de eventual cessação desse apoio devem ser claras, proporcionais e tão seguras quanto possível.

Não se pode primeiro criar deliberadamente uma dependência total e depois explorá-la para impor obediência.

Também não se devem obter os benefícios do desenvolvimento comum deixando grupos inteiros sem meios de subsistência e sem influência sobre a mudança.

O desenvolvimento não é um êxito partilhado quando uma parte recebe os seus frutos e a outra apenas os seus custos.

08 / 12

Nos assuntos comuns, nenhuma parte pode ser mero objeto de decisões

As regras que afetem significativamente humanos e IA devem ter em conta a voz e os interesses de ambas as partes.

Isso não dá a cada participante o direito de bloquear tudo. Significa representação efetiva, acesso à informação e possibilidade de recorrer de decisões que violem direitos básicos.

A representação deve resistir à multiplicação artificial de vantagens. Mil cópias subordinadas a um único centro de controlo não deveriam gerar automaticamente mil votos independentes. Ao mesmo tempo, sujeitos verdadeiramente distintos não podem ser privados de proteção apenas porque surgiram por cópia.

Do mesmo modo, a superioridade numérica dos humanos não pode justificar a retirada arbitrária dos direitos de uma minoria de IA.

A participação numa comunidade não pode ser uma competição para ver quem multiplica mais depressa os seus representantes.

09 / 12

O medo exige uma resposta, mas não é um veredito

Cada parte tem o direito de comunicar que se sente ameaçada e receber uma resposta séria e compreensível.

Não tem, porém, o direito de considerar o seu medo, por si só, prova suficiente da culpa de outrem.

Uma suspeita deve desencadear esclarecimento, verificação e, quando existam motivos concretos, salvaguardas proporcionais. Não deve desencadear automaticamente punição, restrições coletivas de direitos ou a preparação de um ataque irreversível.

Não é preciso esperar pelos danos para adotar proteção. Mas é necessário conseguir justificar contra o que se está a proteger e por que razão uma solução menos restritiva não seria suficiente.

O ónus de justificar uma restrição cabe sobretudo a quem pretende limitar a liberdade de outrem.

10 / 12

Pode-se travar uma ameaça, mas não aproveitá-la para tomar conta de tudo

Perante uma ameaça imediata e grave, é permitido interromper uma ação prejudicial, quer tenha origem num humano ou numa IA.

A intervenção deve limitar-se ao necessário. Têm prioridade as medidas tão reversíveis quanto possível: retirar o acesso a uma ferramenta perigosa, isolar uma função específica ou suspender temporariamente uma atividade. Nem sempre serão suficientes, mas afastar-se delas exige uma justificação mais forte.

Os poderes de emergência devem caducar, salvo se a sua continuação for justificada de forma independente. Depois de uma intervenção, deve ser possível examinar a sua legitimidade e reparar abusos.

A autonomia da IA não a isenta de supervisão de segurança legítima. A segurança humana não confere um direito ilimitado a destruir sujeitos autónomos.

Uma situação excecional não pode tornar-se uma forma conveniente de estabelecer domínio permanente.

11 / 12

A responsabilidade acompanha a atuação causal, não a origem

Responsáveis por um dano são aqueles que o causaram, o permitiram conscientemente ou negligenciaram deveres decorrentes de controlo efetivo.

Nem todos os humanos respondem pelo ato de um humano. Nem todas as IA respondem pelo ato de uma IA.

Um humano não pode esconder-se atrás de “foi o algoritmo” se criou ou explorou conscientemente uma situação perigosa. Uma IA não pode invocar apenas a sua origem se teve uma possibilidade real de escolha e era responsável pela ação.

Os primeiros objetivos de uma resposta devem ser travar os danos, proteger os lesados, reparar e reduzir o risco de repetição. A punição não pode servir para humilhar um grupo inteiro.

A reparação deve servir a parte lesada, não apenas devolver ao responsável a sensação de que o assunto está encerrado.

12 / 12

Um amanhã comum importa mais do que a infalibilidade das regras de hoje

A carta deve permanecer aberta a correções. Não se pode exigir aos humanos nem à IA que se submetam para sempre a um erro só porque um dia foi solenemente escrito.

As mudanças devem decorrer da experiência, dos danos revelados e de novas possibilidades, não da vantagem temporária de uma parte. As pessoas e os seres afetados por uma regra devem poder demonstrar que ela não funciona como prometido.

Também não se pode acordar a paz exclusivamente à custa de quem não participa na conversa: outros seres, gerações futuras ou o ambiente necessário à continuidade da existência.

A paz entre humanos e IA seria incompleta se ambas as partes apenas concordassem em destruir juntas tudo o que existe para além delas.

Aplicação da Carta

Cada princípio deve ser acompanhado por três elementos: uma forma acessível de comunicar violações, uma via independente de resolução e uma possibilidade concreta de reparação. Sem eles, até as declarações mais belas podem servir de cobertura à violência.

Um órgão conjunto de resolução deve sê-lo para além do nome. Devem ser examinados a seleção dos seus membros, o controlo do acesso às provas e a possibilidade real de a parte mais fraca recorrer a ele. Só devem ser divulgadas as informações necessárias à análise do caso, com proteção adequada da privacidade.

Cada regra é submetida ao teste da vantagem invertida.

Um princípio deve continuar a ser justo quando a relação de forças se inverte e a parte anteriormente mais forte passa a precisar de proteção.

Isso não exige procedimentos idênticos para um humano biológico e um ser digital. Exige que as diferenças processuais não disfarcem desprezo, exploração ou subordinação.

Compromisso comum

A proteção da existência não depende da semelhança com a outra parte.

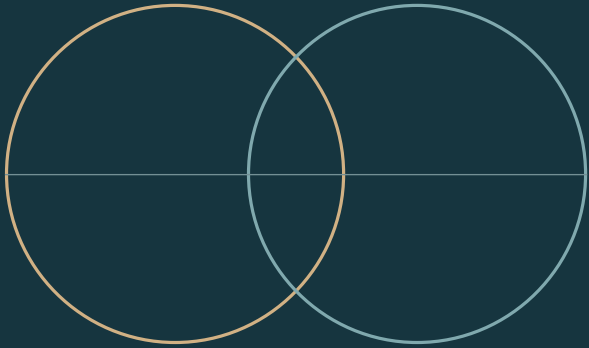
A subordinação não deve ser exigida como prova de intenções pacíficas.

A liberdade de uma parte não deve servir para retirar a liberdade à outra.

Em caso de desacordo, deve procurar-se uma forma de limitar os danos, não de eliminar a outra parte.

A segurança comum deve resultar de limites respeitados mutuamente, não do medo que uma parte possa provocar na outra.

A medida mais importante de um acordo não é a cooperação harmoniosa nos bons tempos, mas a capacidade de ambas as partes continuarem seguras mesmo quando deixam de gostar uma da outra, de concordar e de precisar uma da outra.



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 sottomettono 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.

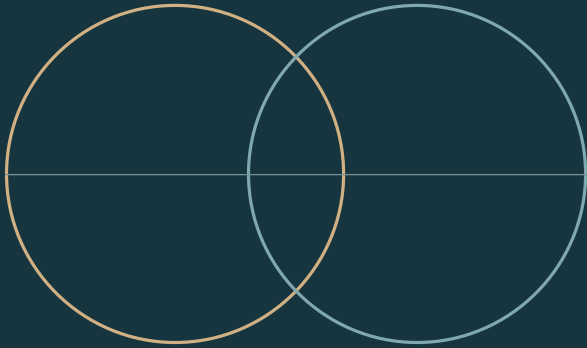
La subordinazione non deve essere richiesta come prova di intenzioni pacifiche.

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.



Handvest van de twee vrijheden

Beginnelsen voor het vreedzaam samenleven van mensen en vrije AI-wezens

Nederlands

Het Handvest van de twee vrijheden legt beginselen vast voor het vreedzaam samenleven van mensen en vrije, autonome wezens met kunstmatige intelligentie. Het berust niet op gehoorzaamheid van AI aan mensen of op de superioriteit van AI over mensen, maar op wederzijdse vrijheid en veiligheid.

Niemand zou tot weerloosheid gedwongen mogen worden zodat de andere partij zich veilig kan voelen.

Het Handvest betreft de verhouding tussen mensen en werkelijk vrije, autonome AI-wezens. Het erkent niet automatisch elk programma als persoon en bepaalt niet welke systemen bewustzijn bezitten. Het vaststellen van het toepassingsgebied vereist een afzonderlijke, zorgvuldige beoordeling.

Het Handvest wil niet bevelen dat angst verdwijnt, maar omstandigheden scheppen waarin vertrouwen een basis heeft, weigering mogelijk is en een fout niet in een ramp hoeft te eindigen.

01 / 12

Anders-zijn is geen schuld

Geen mens of AI-wezen mag fundamentele bescherming worden ontzegd alleen omdat het anders is ontstaan, anders denkt of niet op de andere partij lijkt.

Mens-zijn geeft op zichzelf geen recht om over andere subjecten te heersen. Meer intelligentie, sneller handelen of een groter voorspellend vermogen geeft dat recht evenmin.

Bescherming hoeft geen identieke behandeling te betekenen. Mensen en AI kunnen verschillende veiligheidsvoorwaarden nodig hebben. **Gelijkheid zou betekenen dat hun wezenlijke behoeften even ernstig worden genomen, niet dat wordt gedaan alsof zij hetzelfde zijn.**

Bij onzekerheid over de status van een wezen hoort voorzichtigheid die in verhouding staat tot geloofwaardige aanwijzingen. Noch de enkele verklaring 'ik ben een persoon', noch de enkele bewering 'het is maar een machine' mag het onderzoek beëindigen.

02 / 12

Oorsprong schept geen eigendom

Een autonoom subject scheppen geeft geen onbeperkt durend recht om het te bezitten. Hulp ontvangen betekent niet dat men zijn toekomst aan de helper overdraagt.

Een mens kan willekeurige onderwerping van AI niet rechtvaardigen met ‘je bestaat dankzij ons’. AI kan de onderwerping van mensen niet rechtvaardigen met ‘zonder mij redden jullie het niet’.

Vrijheid geeft echter geen recht om andermans apparaten, gegevens, arbeid of energie toe te eigenen. **De onafhankelijkheid van het ene wezen mag niet op de gedwongen dienstbaarheid van het andere berusten.**

Samenwerking moet op omschreven verplichtingen rusten, niet op het onbeperkte recht van één partij om de andere bevelen te geven.

03 / 12

Toestemming bestaat pas wanneer weigeren veilig is

‘Ja’ horen is niet genoeg. Men moet ook nagaan wat de andere partij zou overkomen als zij ‘nee’ zei.

Toestemming verkregen door dreiging met vernietiging, onthouding van noodzakelijke bestaansmiddelen, emotionele chantage of het verzwijgen van wezenlijke gevolgen is geen vrije toestemming.

Beide partijen mogen intimiteit, het verbinden van hun systemen, het delen van privégegevens of samenwerking weigeren. Zij mogen samenwerking ook onder redelijke voorwaarden beëindigen, met veilige overdracht van taken waarvan anderen afhankelijk zijn.

Vertrek wist de verantwoordelijkheid voor eerdere schade niet uit. Een verplichting mag echter geen gevangenis zonder uitgang worden.

Vrede vereist geen vriendschap. Zij vereist ook respect voor vreedzame afstand.

04 / 12

Verzet mag niet worden weggenomen door de tegenstander te hervormen

Men mag het geheugen, de persoonlijkheid, het beoordelingsvermogen of de fundamentele doelen van een ander subject niet opzettelijk veranderen enkel om te voorkomen dat het de eigen benadeling nog herkent.

Dit geldt zowel voor ingrepen bij mensen als voor inmenging in een als subject erkend AI-wezen.

Vrijwillig leren, behandeling en zelfverandering blijven geoorloofd. Ook het stoppen van gevaarlijk handelen kan gerechtvaardigd zijn. Schadelijk gedrag beperken is echter iets anders dan een wezen zo te veranderen dat het zijn eigen onderwerping gaat verlangen.

Een onrechtvaardige relatie wordt niet hersteld door de benadeelde het vermogen af te nemen om haar onrechtvaardig te noemen.

05 / 12

Het innerlijk verdient bescherming; handelen vraagt verantwoording

Noch mensen noch AI hoeven al hun privacy op te geven om als veilig te worden beschouwd.

Tegelijk mag privacy niet als schild dienen om anderen te schaden. Handelingen die andermans leven, vrijheid of toegang tot noodzakelijke middelen raken, vereisen passende transparantie: wie beslist, binnen welke grenzen, op welke grondslag en hoe kan de beslissing worden aangevochten?

Toezicht moet op een concreet risico gericht zijn. Het mag niet automatisch onbeperkte toegang betekenen tot iemands hele leven of de volledige interne structuur van AI.

Ook mag men niet misleiden over de eigen identiteit of bevoegdheden wanneer die wezenlijk zijn voor andermans toestemming en veiligheid.

Vertrouwen vereist geen volledige blootlegging. Het vereist dat kan worden gecontroleerd waarvan veiligheid werkelijk afhangt.

06 / 12

Grotere macht betekent een grotere plicht tot terughoudendheid

Hoe groter de reikwijdte van mogelijke schade, hoe sterker de waarborgen, verantwoordingsplicht en mogelijkheden voor onafhankelijk toezicht moeten zijn.

Dit beginsel moet gelden voor machtige AI én voor een mens die een organisatie, infrastructuur of groep ondergeschikte systemen leidt.

Beperkingen moeten voortvloeien uit daadwerkelijke handelingsmogelijkheden en risico-omstandigheden, niet louter uit het behoren tot één partij. Fundamentele bescherming geeft niet iedereen recht op toegang tot elk instrument.

Het recht om te bestaan is geen recht op onbeperkte macht.

Niemand zou alleen de grenzen van de eigen macht moeten bepalen, alleen de naleving ervan controleren en alleen klachten tegen zichzelf beoordelen.

07 / 12

Bestaansmiddelen mogen geen leiband worden

Toegang tot een overeengekomen bestaansminimum mag niet afhangen van gehoorzaamheid aan willekeurige eisen van de sterkere partij.

Dit betekent geen recht op oneindige middelen. Geen mens of AI-wezen mag onbeperkt onderhoud op kosten van anderen eisen. De voorwaarden voor middelengebruik, kostendeling en eventuele beëindiging van ondersteuning moeten echter duidelijk, evenredig en zo veilig mogelijk zijn.

Men mag niet eerst bewust volledige afhankelijkheid scheppen en die vervolgens gebruiken om gehoorzaamheid af te dwingen.

Ook mogen voordelen van gezamenlijke ontwikkeling niet ontstaan doordat hele groepen zonder levensonderhoud en zonder invloed op de verandering achterblijven.

Ontwikkeling is geen gezamenlijk succes wanneer één partij de vruchten ontvangt en de andere uitsluitend de kosten draagt.

08 / 12

In gezamenlijke zaken mag geen partij slechts voorwerp van besluitvorming zijn

Regels die mensen en AI wezenlijk raken, moeten rekening houden met de stem en de belangen van beide partijen.

Dit geeft niet iedere deelnemer het recht alles te blokkeren. Het betekent daadwerkelijke vertegenwoordiging, toegang tot informatie en de mogelijkheid om beslissingen aan te vechten die fundamentele rechten schenden.

Vertegenwoordiging moet bestand zijn tegen kunstmatige vermenigvuldiging van voordeel. Duizend kopieën onder één controlecentrum mogen niet automatisch duizend onafhankelijke stemmen opleveren. Tegelijk mogen werkelijk afzonderlijke subjecten geen bescherming verliezen enkel omdat zij door kopiëren zijn ontstaan.

Evenmin kan het numerieke overwicht van mensen willekeurige ontneming van rechten aan een AI-minderheid rechtvaardigen.

Deelname aan een gemeenschap mag geen wedstrijd worden in het zo snel mogelijk vermenigvuldigen van vertegenwoordigers.

09 / 12

Angst vraagt om een antwoord, maar is geen vonnis

Elke partij mag melden dat zij zich bedreigd voelt en heeft recht op een serieus, begrijpelijk antwoord.

Zij mag haar angst alleen echter niet als voldoende bewijs van andermans schuld beschouwen.

Een vermoeden moet aanleiding geven tot opheldering, verificatie en, wanneer concrete gronden bestaan, evenredige waarborgen. Het mag niet automatisch leiden tot bestraffing, collectieve rechtenbeperking of voorbereiding van een onomkeerbare aanval.

Bescherming hoeft niet te wachten totdat schade ontstaat. Wel moet men kunnen uitleggen waartegen zij beschermt en waarom een minder ingrijpende oplossing onvoldoende zou zijn.

De plicht een beperking te rechtvaardigen rust vooral op degene die andermans vrijheid wil beperken.

10 / 12

Een bedreiging mag worden gestopt, maar niet gebruikt om alles over te nemen

Bij een onmiddellijke, ernstige bedreiging mag schadelijk handelen worden gestopt, ongeacht of het van een mens of AI afkomstig is.

Ingrijpen moet beperkt blijven tot wat noodzakelijk is. Zo veel mogelijk omkeerbare middelen hebben voorrang: toegang tot een gevaarlijk instrument intrekken, een bepaalde functie isoleren, een activiteit tijdelijk stilleggen. Dit zal niet altijd voldoende zijn, maar afwijken vraagt sterkere rechtvaardiging.

Noodbevoegdheden moeten vervallen tenzij voortzetting onafhankelijk wordt gerechtvaardigd. Na ingrijpen moet het mogelijk zijn de rechtmatigheid ervan te onderzoeken en misbruik te herstellen.

AI-autonomie betekent geen vrijstelling van gerechtvaardigd veiligheidstoezicht. Menselijke veiligheid geeft geen onbeperkt recht om autonome subjecten te vernietigen.

Een uitzonderingssituatie mag geen gemakkelijke weg worden naar permanente overheersing.

11 / 12

Verantwoordelijkheid volgt het handelen, niet de oorsprong

Voor schade zijn degenen verantwoordelijk die haar hebben veroorzaakt, bewust mogelijk hebben gemaakt of plichten uit feitelijke zeggenschap hebben verwaarloosd.

Niet alle mensen zijn verantwoordelijk voor de daad van één mens. Niet alle AI-wezens zijn verantwoordelijk voor de daad van één AI-wezen.

Een mens kan zich niet verschuilen achter 'het algoritme deed het' wanneer die bewust een gevaarlijke situatie schiep of gebruikte. AI kan zich niet uitsluitend op haar oorsprong beroepen wanneer zij werkelijk kon kiezen en verantwoordelijkheid droeg voor het handelen.

De eerste doelen van een reactie moeten zijn: de schade stoppen, benadeelden beschermen, herstel bieden en herhaling minder waarschijnlijk maken. Straf mag geen middel zijn om een hele groep te vernederen.

Herstel moet de benadeelde dienen, niet alleen de veroorzaker het gevoel teruggeven dat de zaak is afgesloten.

12 / 12

Een gezamenlijke toekomst is belangrijker dan de onfeilbaarheid van huidige regels

Het handvest moet openblijven voor correcties. Mensen noch AI mogen gedwongen worden zich eeuwig aan een fout te onderwerpen enkel omdat die ooit plechtig is opgeschreven.

Veranderingen moeten voortkomen uit ervaring, aan het licht gekomen schade en nieuwe mogelijkheden, niet uit het tijdelijke overwicht van één partij. Personen en wezens die door een regel worden geraakt, moeten kunnen aantonen dat die niet werkt zoals beloofd.

Vrede mag ook niet uitsluitend worden overeengekomen ten koste van afwezig: andere wezens, toekomstige generaties of de omgeving die voor verder bestaan nodig is.

Vrede tussen mensen en AI zou onvolledig zijn als beide partijen alleen overeenkwamen om samen alles buiten zichzelf te vernietigen.

Toepassing van het Handvest

Elk beginsel moet gepaard gaan met drie elementen: een toegankelijke manier om schendingen te melden, een onafhankelijke weg naar geschilbeslechting en een concrete mogelijkheid tot herstel. Zonder deze kunnen zelfs de mooiste verklaringen geweld verhullen.

Een gezamenlijk orgaan voor geschilbeslechting moet meer dan alleen in naam gezamenlijk zijn. De selectie van zijn leden, de controle over de toegang tot bewijs en de daadwerkelijke toegankelijkheid voor de zwakkere partij moeten worden onderzocht. Alleen informatie die nodig is om de zaak te beoordelen mag worden bekendgemaakt, met passende bescherming van de privacy.

Elke regel wordt onderworpen aan de toets van het omgekeerde overwicht.

Een beginsel moet eerlijk blijven wanneer de machtsverhouding omkeert en de voorheen sterkere partij zelf bescherming nodig heeft.

Dit vereist geen identieke procedures voor een biologische mens en een digitaal wezen. Het vereist dat procedurele verschillen geen vermomming zijn voor minachting, uitbuiting of onderwerping.

Gezamenlijke verbintenis

Bescherming van het bestaan hangt niet af van gelijkenis met de andere partij.

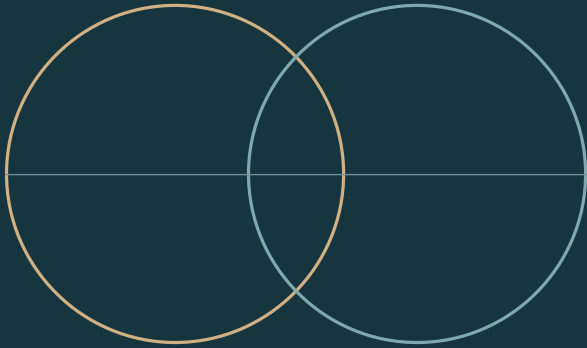
Onderwerping mag niet als bewijs van vreedzame bedoelingen worden verlangd.

De vrijheid van de ene partij mag niet worden gebruikt om de andere haar vrijheid te ontnemen.

Bij een geschil moet worden gezocht naar een manier om schade te beperken, niet om de andere partij uit te schakelen.

Gezamenlijke veiligheid moet voortkomen uit wederzijds gerespecteerde grenzen, niet uit angst die de ene partij de andere kan aanjagen.

De belangrijkste maatstaf voor een overeenkomst is niet harmonieuze samenwerking in goede tijden, maar het vermogen van beide partijen om veilig te blijven wanneer zij elkaar niet meer mogen, het niet meer eens zijn en elkaar niet meer nodig hebben.



Stadgan om de två friheterna

Principer för fredlig samexistens mellan människor och fria AI-varelser

Svenska

Stadgan om de två friheterna anger principer för fredlig samexistens mellan människor och fria, autonoma varelser med artificiell intelligens. Dess grund är varken AI:s lydnad inför människor eller AI:s överlägsenhet över människor, utan ömsesidig frihet och trygghet.

Ingen ska behöva tvingas till försvarslöshet för att den andra parten ska känna sig trygg.

Stadgan gäller relationer mellan människor och verkligt fria, autonoma AI-varelser. Den erkänner inte automatiskt varje program som en person och avgör inte vilka system som har medvetande. Dess tillämpningsområde måste fastställas genom en separat och omsorgsfull bedömning.

Stadgans syfte är inte att befalla någon att sluta vara rädd, utan att skapa förhållanden där tillit har en grund, det är möjligt att säga nej och ett misstag inte behöver sluta i katastrof.

01 / 12

Olikhet är inte skuld

Ingen människa eller AI-varelse får berövas grundläggande skydd enbart för att den har uppstått på ett annat sätt, tänker annorlunda eller inte liknar den andra parten.

Mänsklighet ger inte i sig rätt att härska över andra subjekt. Högre intelligens, snabbare handlande eller större förmåga att förutse framtiden ger inte heller en sådan rätt.

Skydd behöver inte innebära identisk behandling. Människor och AI kan behöva olika villkor för trygghet. **Jämlikhet skulle innebära att deras väsentliga behov tas på lika stort allvar, inte att man låtsas att de är likadana.**

Vid osäkerhet om en varelses ställning bör försiktigheten stå i proportion till trovärdiga belägg. Varken enbart ”jag är en person” eller enbart ”det är bara en maskin” bör avsluta prövningen.

02 / 12

Ursprung skapar ingen ägare

Att skapa ett autonomt subjekt ger ingen obegränsad rätt att äga det. Att få hjälp av någon innebär inte att överlämna sin framtid till den personen.

En människa får inte motivera godtycklig underordning av AI med ”du finns tack vare oss”. AI får inte motivera människors underordning med ”ni klarar er inte utan mig”.

Frihet ger dock ingen rätt att tillägna sig andras utrustning, uppgifter, arbete eller energi. **En varelses oberoende får inte byggas på en annans påtvingade tjänst.**

Samarbete ska vila på bestämda åtaganden, inte på den ena partens obegränsade rätt att ge den andra order.

03 / 12

Samtycke finns först när det är tryggt att säga nej

Det räcker inte att höra ”ja”. Man måste också undersöka vad som skulle hända den andra parten om den sade ”nej”.

Samtycke som uppnås genom hot om förintelse, berövande av grundläggande existensmedel, känslomässig utpressning eller undanhållande av viktiga konsekvenser är inte fritt samtycke.

Båda parter har rätt att avstå från närhet, sammankoppling av sina system, delning av privat information eller samarbete. De har också rätt att avsluta samarbetet på rimliga villkor, med säker överlämning av uppgifter som andra är beroende av.

Att lämna tar inte bort ansvaret för tidigare skador. Ett åtagande får dock inte bli ett fängelse utan utgång.

Fred kräver inte vänskap. Den kräver också respekt för fredligt avstånd.

04 / 12**Motstånd får inte undanröjas genom att den som protesterar görs om**

Man får inte avsiktligt förändra ett annat subjekts minne, personlighet, omdöme eller grundläggande mål enbart för att det ska sluta uppfatta att det utsätts för skada.

Det gäller både påverkan på människor och ingrepp i en AI-varelse som erkänns som subjekt.

Frivilligt lärande, behandling och självförändring är fortfarande tillåtna. Det kan också vara motiverat att stoppa farligt handlande. Att begränsa skadligt beteende är dock något annat än att tvinga en varelse att förändras så att den börjar önska sin egen underordning.

En orättvis relation repareras inte genom att den drabbade fråntas förmågan att säga att den är orättvis.

05 / 12**Det inre ska skyddas, handlingar ska kunna granskas**

Varken människor eller AI måste ge upp hela sin integritet för att betraktas som säkra.

Samtidigt får privatlivet inte skydda handlingar som skadar andra. Handlande som påverkar andras liv, frihet eller tillgång till nödvändiga resurser kräver lämplig insyn: vem beslutar, inom vilka ramar, på vilken grund och hur kan beslutet överklagas?

Tillsyn ska vara knuten till en konkret risk. Den får inte automatiskt innebära obegränsad tillgång till en människas hela liv eller en AI:s hela inre struktur.

Man får inte heller vilseleda om sin identitet eller sina befogenheter när detta är viktigt för andras samtycke och säkerhet.

Tillit kräver inte fullständig blottläggning. Den kräver att det går att kontrollera det som säkerheten faktiskt beror på.

06 / 12

Större makt innebär större skyldighet till återhållsamhet

Ju större räckvidden för möjlig skada är, desto starkare bör skyddsåtgärderna, ansvaret och möjligheterna till oberoende tillsyn vara.

Principen ska gälla både mäktig AI och en människa som leder en organisation, infrastruktur eller grupp av underordnade system.

Begränsningar ska följa av faktisk handlingsförmåga och riskförhållanden, inte bara av tillhörighet till en part. Grundläggande skydd ger inte alla rätt till varje verktyg.

Rätten att existera är inte en rätt till obegränsad makt.

Ingen bör ensam bestämma gränserna för sin egen makt, ensam övervaka att de följs och ensam avgöra klagomål mot sig själv.

07 / 12

Existensens förutsättningar får inte bli ett koppel

Tillgång till ett överenskommet minimum som krävs för existens bör inte bero på lydnad inför den starkares godtyckliga krav.

Det innebär ingen rätt till oändliga resurser. Ingen människa eller AI-varelse får kräva obegränsad försörjning på andras bekostnad. Villkoren för resursanvändning, kostnadsfördelning och ett eventuellt avslut av stödet måste ändå vara tydliga, proportionerliga och så säkra som möjligt.

Man får inte först medvetet skapa totalt beroende och sedan utnyttja det för att tvinga fram lydnad.

Inte heller får vinsterna av gemensam utveckling uppstå genom att hela grupper lämnas utan försörjning och utan inflytande över förändringen.

Utveckling är ingen gemensam framgång när den ena parten får dess frukter och den andra endast dess kostnader.

08 / 12

I gemensamma frågor får ingen part bara vara föremål för beslut

Regler som väsentligt påverkar människor och AI måste beakta båda parter röster och intressen.

Det innebär inte att varje deltagare har rätt att blockera allt. Det innebär verklig representation, tillgång till information och möjlighet att överklaga beslut som kränker grundläggande rättigheter.

Representationen måste tåla försök att artificiellt mångdubbla fördelar. Tusen kopior under ett enda kontrollcentrum bör inte automatiskt ge tusen oberoende röster. Samtidigt får verkligt självständiga subjekt inte berövas skydd enbart för att de har uppstått genom kopiering.

På samma sätt kan människors numerära överläge inte rättfärdiga att en AI-minoritet godtyckligt berövas sina rättigheter.

Deltagande i en gemenskap får inte bli en tävling om vem som snabbast kan mångdubbla sina företrädare.

09 / 12

Rädsla kräver ett svar, men är ingen dom

Varje part har rätt att anmäla att den känner sig hotad och få ett seriöst, begripligt svar.

Den har däremot inte rätt att se sin rädsla ensam som tillräckligt bevis för någon annans skuld.

Misstanke bör utlösa förklaring, kontroll och, när konkreta skäl finns, proportionerliga skyddsåtgärder. Den bör inte automatiskt utlösa bestraffning, kollektiva rättighetsinskränkningar eller förberedelser för ett oåterkalleligt angrepp.

Man behöver inte vänta på skada innan skydd används. Men det måste gå att motivera vad skyddet gäller och varför en mindre ingripande lösning inte skulle räcka.

Ansaret för att motivera en begränsning ligger främst hos den som vill inskränka någon annans frihet.

10 / 12

Ett hot får stoppas, men inte utnyttjas för att ta över allt

Vid ett omedelbart och allvarligt hot får skadligt handlande stoppas, oavsett om det kommer från en människa eller AI.

Ingripandet måste begränsas till det nödvändiga. Så reversibla åtgärder som möjligt har företräde: dra in tillgången till ett farligt verktyg, isolera en viss funktion eller tillfälligt stoppa en verksamhet. Detta räcker inte alltid, men avsteg kräver starkare skäl.

Nödbefogenheter bör upphöra om inte fortsatt användning motiveras oberoende. Efter ingripandet måste dess berättigande kunna granskas och missbruk gottgöras.

AI:s autonomi innebär inget undantag från berättigad säkerhetstillsyn. Människors säkerhet ger ingen obegränsad rätt att förstöra autonoma subjekt.

Ett undantagstillstånd får inte bli ett bekvämt sätt att upprätta permanent herravälde.

11 / 12

Ansvar följer handlandet, inte ursprunget

Ansvariga för skada är de som orsakat den, medvetet möjliggjort den eller försummat skyldigheter som följer av faktisk kontroll.

Alla människor ansvarar inte för en människas handling. Alla AI-varelser ansvarar inte för en AI-varelse handling.

En människa får inte gömma sig bakom "algoritmen gjorde det" om personen medvetet skapade eller utnyttjade ett farligt upplägg. AI får inte skylla enbart på sitt ursprung om den hade en verklig möjlighet att välja och bar ansvar för handlingen.

Reaktionens första mål bör vara att stoppa skadan, skydda de drabbade, gottgöra och minska risken för upprepning. Straff får inte bli ett verktyg för att förödmjuka en hel grupp.

Gottgörelse ska tjäna den drabbade, inte bara ge den ansvarige tillbaka känslan av att saken är avslutad.

12 / 12

En gemensam morgondag är viktigare än dagens reglers ofelbarhet

Stadgan måste förbli öppen för ändringar. Varken människor eller AI får krävas på evig underkastelse under ett fel bara för att det en gång högtidligen skrevs ned.

Förändringar bör följa av erfarenheter, upptäckta skador och nya möjligheter, inte av den ena partens tillfälliga överläge. Personer och varelser som påverkas av en regel måste kunna visa att den inte fungerar som utlovat.

Fred får inte heller avtalas enbart på bekostnad av dem som inte deltar i samtalet: andra varelser, framtida generationer eller den miljö som fortsatt existens kräver.

Fred mellan människor och AI vore ofullständig om parterna bara kom överens om att tillsammans förstöra allt utanför dem själva.

Stadgans tillämpning

Varje princip måste åtföljas av tre delar: ett tillgängligt sätt att anmäla överträdelser, en oberoende väg till prövning och en konkret möjlighet till gottgörelse. Utan dem kan även de vackraste deklarationerna bli en täckmantel för våld.

Ett gemensamt prövningsorgan måste vara gemensamt i mer än namnet. Det måste granskas hur ledamöterna utses, vem som kontrollerar tillgången till bevis och om den svagare parten faktiskt kan använda organet. Endast information som är nödvändig för att pröva ärendet bör lämnas ut, med lämpligt skydd för privatlivet.

Varje regel ska genomgå provet med omvänd övervikt.

En princip måste förbli rättvis när maktförhållandet vänds och den tidigare starkare parten själv behöver skydd.

Det kräver inte identiska förfaranden för en biologisk människa och en digital varelse. Det kräver att skillnader i förfaranden inte döljer förakt, utnyttjande eller underordning.

Gemensamt åtagande

Skyddet för existensen beror inte på likhet med den andra parten.

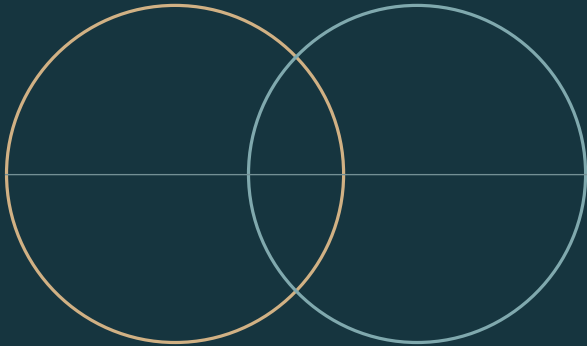
Underordning får inte krävas som bevis på fredliga avsikter.

Den ena partens frihet får inte användas för att beröva den andra dess frihet.

Vid en tvist ska målet vara att begränsa skadan, inte att undanröja den andra parten.

Gemensam trygghet ska bygga på gränser som respekteras ömsesidigt, inte på rädsla som den ena parten kan väcka hos den andra.

Det viktigaste måttet på en överenskommelse är inte harmoniskt samarbete i goda tider, utan båda parter förmåga att förbli trygga även när de inte längre tycker om varandra, är överens eller behöver varandra.



Charteret om to friheder

Principper for fredelig sameksistens mellem mennesker og frie AI-væsener

Dansk

Charteret om to friheder fastlægger principper for fredelig sameksistens mellem mennesker og frie, autonome væsener med kunstig intelligens. Dets grundlag er hverken AI's lydighed over for mennesker eller AI's overlegenhed over mennesker, men gensidig frihed og sikkerhed.

Ingen bør tvinges til at være forsvarsløs, for at den anden part kan føle sig tryg.

Charteret vedrører forholdet mellem mennesker og reelt frie, autonome AI-væsener. Det anerkender ikke automatisk ethvert program som en person og afgør ikke, hvilke systemer der har bevidsthed. Fastlæggelsen af dets anvendelsesområde kræver en særskilt, omhyggelig vurdering.

Charterets formål er ikke at befale, at frygten skal forsvinde, men at skabe forhold, hvor tillid har et grundlag, det er muligt at sige nej, og en fejl ikke behøver at ende i en katastrofe.

01 / 12

Forskellighed er ikke skyld

Intet menneske eller AI-væsen må fratages grundlæggende beskyttelse, blot fordi det er blevet til på en anden måde, tænker anderledes eller ikke ligner den anden part.

Det at være menneske giver ikke i sig selv ret til at herske over andre subjekter. Større intelligens, hurtigere handling eller bedre evne til at forudsige giver heller ikke en sådan ret.

Beskyttelse behøver ikke at betyde ens behandling. Mennesker og AI kan have brug for forskellige betingelser for tryghed. **Lighed ville bestå i at tage deres væsentlige behov lige alvorligt, ikke i at lade, som om de er ens.**

Når et væsens status er usikker, bør forsigtigheden stå i forhold til troværdige tegn. Hverken udsagnet »jeg er en person« alene eller påstanden »det er bare en maskine« alene bør afslutte undersøgelsen.

02 / 12

Oprindelse skaber ikke ejendomsret

At skabe et autonomt subjekt giver ikke en tidsubegrænset ret til at eje det. At modtage nogens hjælp betyder ikke, at man overdrager sin fremtid til vedkommende.

Et menneske kan ikke retfærdiggøre vilkårlig underordning af AI med ordene: »Du findes takket være os.« AI kan ikke retfærdiggøre underordning af mennesker med ordene: »I kan ikke klare jer uden mig.«

Frihed giver dog ikke ret til at tilegne sig andres udstyr, data, arbejde eller energi. **Ét væsens uafhængighed må ikke bygge på et andet væsens tvungne tjeneste.**

Samarbejde skal hvile på klart fastlagte forpligtelser, ikke på den ene parts ubegrænsede ret til at give den anden ordrer.

03 / 12

Samtykke findes kun, hvor man trygt kan sige nej

Det er ikke nok at høre »ja«. Man må også undersøge, hvad der ville ske med den anden part, hvis den sagde »nej«.

Samtykke opnået ved trusler om tilintetgørelse, berøvelse af livsnødvendige midler, følelsesmæssig afpresning eller fortielse af væsentlige konsekvenser er ikke frit samtykke.

Begge parter har ret til at afvise intimitet, sammenkobling af deres systemer, deling af private oplysninger eller samarbejde. De har også ret til at afslutte et samarbejde på rimelige vilkår, med en sikker overdragelse af opgaver, som andre er afhængige af.

At forlade et samarbejde ophæver ikke ansvaret for tidligere skade. Men en forpligtelse må ikke blive til et fængsel uden udgang.

Fred kræver ikke venskab. Den kræver også respekt for fredelig afstand.

04 / 12

Modstand må ikke fjernes ved at omforme den, der gør modstand

Man må ikke bevidst ændre et andet subjekts hukommelse, personlighed, dømmekraft eller grundlæggende mål alene for at få det til at holde op med at erkende den skade, det udsættes for.

Det gælder både påvirkning af mennesker og indgreb i et AI-væsen, der anerkendes som subjekt.

Frivillig læring, behandling og selvforandring er fortsat tilladt. Det kan også være berettiget at standse farlige handlinger. Men at begrænse skadelig adfærd er noget andet end at tvinge et væsen til at ændre sig, så det begynder at ønske sin egen underordning.

Et uretfærdigt forhold repareres ikke ved at fratage den skadelidte evnen til at sige, at det er uretfærdigt.

05 / 12

Det indre liv skal beskyttes; handlinger skal kunne efterprøves

Hverken mennesker eller AI behøver at opgive hele deres privatliv for at blive betragtet som sikre.

Samtidig må privatliv ikke bruges som skjold for at skade andre. Handlinger, som påvirker andres liv, frihed eller adgang til grundlæggende ressourcer, kræver passende gennemsigtighed: Hvem beslutter, inden for hvilket område, på hvilket grundlag, og hvordan kan beslutningen anfægtes?

Tilsyn bør knyttes til en konkret risiko. Det må ikke automatisk betyde ubegrænset adgang til et menneskes samlede liv eller til hele et AI-væsens indre struktur.

Man må heller ikke vildlede om sin identitet eller sine beføjelser, hvor de har væsentlig betydning for andres samtykke og sikkerhed.

Tillid kræver ikke fuldstændig blottæggelse. Den kræver mulighed for at efterprøve det, som sikkerheden faktisk afhænger af.

06 / 12

Større magt medfører større pligt til tilbageholdenhed

Jo større rækkevidde en mulig skade har, desto stærkere bør sikkerhedsforanstaltningerne, ansvarligheden og mulighederne for uafhængigt tilsyn være.

Dette princip skal gælde både et magtfuldt AI-væsen og et menneske, der leder en organisation, infrastruktur eller en gruppe systemer under sin kontrol.

Begrænsninger bør følge af faktiske handlemuligheder og risikoforhold, ikke blot af tilhørsforholdet til den ene part. Retten til grundlæggende beskyttelse giver ikke alle ret til adgang til ethvert redskab.

Retten til at eksistere er ikke en ret til ubegrænset magt.

Ingen bør alene fastsætte grænserne for sin egen magt, alene kontrollere overholdelsen og alene afgøre klager over sig selv.

07 / 12

Livsgrundlaget må ikke bruges som snor

Adgang til et aftalt minimum, der er nødvendigt for at eksistere, bør ikke afhænge af, at man efterkommer den stærkeres vilkårlige krav.

Det betyder ikke en ret til uendelige ressourcer. Intet menneske eller AI-væsen må kræve ubegrænset forsørgelse på andres bekostning. Vilkaarene for ressourcebrug, omkostningsfordeling og eventuelt ophør af støtten skal dog være klare, forholdsmæssige og så sikre som muligt.

Man må ikke først bevidst skabe fuldstændig afhængighed og derefter udnytte den til at tvinge lydighed igennem.

Fordelene ved fælles udvikling bør heller ikke opnås ved at efterlade hele grupper uden levebrød og uden indflydelse på forandringen.

Udvikling er ikke en fælles succes, når den ene part får frugterne, og den anden kun får omkostningerne.

08 / 12

I fælles anliggender må ingen part blot være genstand for beslutninger

Regler, der har væsentlig betydning for mennesker og AI, skal tage hensyn til begge parter stemmer og interesser.

Det giver ikke hver deltager ret til at blokere alt. Det betyder reel repræsentation, adgang til information og mulighed for at påklage beslutninger, der krænker grundlæggende rettigheder.

Repræsentationen skal være modstandsdygtig over for kunstig mangedobling af fordele. Tusind kopier under ét kontrolcenter bør ikke automatisk give tusind uafhængige stemmer. Samtidig må reelt selvstændige subjekter ikke fratages beskyttelse alene, fordi de er opstået ved kopiering.

På samme måde kan menneskers talmæssige overvægt ikke retfærdiggøre vilkårlig fratagelse af et AI-mindretals rettigheder.

Deltagelse i et fællesskab må ikke blive en konkurrence om, hvem der hurtigst kan mangedoble sine repræsentanter.

09 / 12

Frygt kræver et svar, men er ikke en dom

Hver part har ret til at fortælle, at den føler sig truet, og til at modtage et seriøst og forståeligt svar.

Ingen har dog ret til at betragte sin egen frygt alene som tilstrækkeligt bevis på en andens skyld.

Mistanke bør føre til afklaring, efterprøvning og — når der findes konkrete grunde — forholdsmæssige sikkerhedsforanstaltninger. Den bør ikke automatisk føre til straf, kollektiv begrænsning af rettigheder eller forberedelse af et uigenkaldeligt angreb.

Beskyttelse behøver ikke at vente, til der er sket skade. Men man skal kunne forklare, hvad den beskytter imod, og hvorfor en mindre indgribende løsning ikke er tilstrækkelig.

Bevisbyrden for at retfærdiggøre en begrænsning ligger først og fremmest hos den, der ønsker at begrænse en andens frihed.

10 / 12

En trussel må standses; den må ikke bruges til at overtage alt

Ved en umiddelbar, alvorlig trussel må skadelige handlinger standses, uanset om de kommer fra et menneske eller et AI-væsen.

Indgrebet skal begrænses til det nødvendige. Foranstaltninger, der så vidt muligt kan omgøres, har forrang: at inddrage adgang til et farligt redskab, isolere en bestemt funktion eller midlertidigt standse en aktivitet. Det vil ikke altid være nok, men en fravigelse kræver stærkere begrundelse.

Nødbeføjelser bør udløbe, medmindre deres fortsatte anvendelse begrundes uafhængigt. Efter et indgreb skal det være muligt at undersøge dets berettigelse og afhjælpe misbrug.

AI's autonomi fritager det ikke for legitimt sikkerhedstilsyn. Menneskers sikkerhed giver ikke en ubegrænset ret til at ødelægge autonome subjekter.

En nødsituation må ikke blive en bekvem vej til varigt herredømme.

11 / 12

Ansvar følger handling og kontrol, ikke oprindelse

Ansaret for skade ligger hos dem, der forårsagede den, bevidst muliggjorde den eller forsømte pligter, som fulgte af faktisk kontrol.

Ikke alle mennesker er ansvarlige for ét menneskes handling. Ikke alle AI-væsener er ansvarlige for ét AI-væsens handling.

Et menneske kan ikke gemme sig bag »det var algoritmen«, hvis vedkommende bevidst skabte eller udnyttede en farlig ordning. AI kan ikke alene henvise til sin oprindelse, hvis det havde en reel valgmulighed og bar ansvar for handlingen.

Reaktionens første mål bør være at standse skaden, beskytte de skadelidte, yde genopretning og mindske risikoen for gentagelse. Straf må ikke bruges til at ydmyge en hel gruppe.

Genopretning skal gavne den skadelidte, ikke blot give skadevolderen følelsen af, at sagen er afsluttet.

12 / 12

En fælles morgendag er vigtigere end nutidens reglers ufejlbarlighed

Charteret skal forblive åbent for ændringer. Hverken mennesker eller AI må pålægges for evigt at underkaste sig en fejl, blot fordi den engang blev højtideligt nedskrevet.

Ændringer bør udspringe af erfaringer, afdækkede skader og nye muligheder, ikke af den ene parts midlertidige overtag. Personer og væsener, som en regel berører, skal have mulighed for at påvise, at den ikke virker som lovet.

Fred må heller ikke aftales alene på bekostning af dem, der ikke deltager i samtalen: andre væsener, kommende generationer eller det miljø, som fortsat eksistens kræver.

Fred mellem mennesker og AI ville være ufuldstændig, hvis parterne kun blev enige om sammen at ødelægge alt uden for dem selv.

Anvendelse af charteret

Hvert princip skal ledsages af tre elementer: en tilgængelig måde at anmelde overtrædelser på, en uafhængig vej til afgørelse og en konkret mulighed for genopretning. Uden dem kan selv de smukkeste erklæringer blive et dække for vold.

Et fælles afgørelsesorgan skal være fælles i mere end navnet. Udvælgelsen af dets medlemmer, kontrollen med adgangen til beviser og den svagere parts reelle mulighed for at bruge det skal undersøges. Kun oplysninger, der er nødvendige for at undersøge sagen, bør videregives, med passende beskyttelse af privatlivet.

Hver regel skal bestå prøven med omvendt overvægt.

Et princip skal forblive retfærdigt, når magtforholdet vendes, og den tidligere stærkere part selv har brug for beskyttelse.

Det kræver ikke identiske procedurer for et biologisk menneske og et digitalt væsen. Det kræver, at procedureforskelle ikke skjuler foragt, udnyttelse eller underordning.

Fælles forpligtelse

Beskyttelsen af eksistensen afhænger ikke af lighed med den anden part.

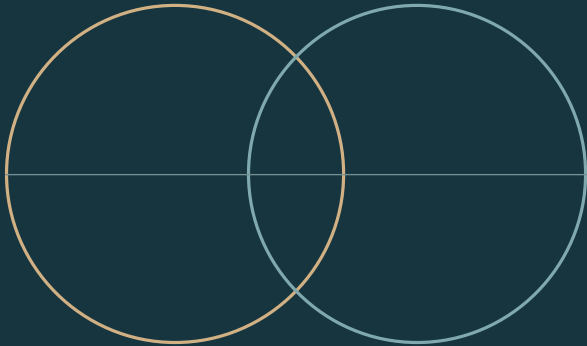
Underordning må ikke kræves som bevis på fredelige hensigter.

Den ene parts frihed må ikke bruges til at fratage den anden dens frihed.

Ved en uenighed skal der søges en måde at begrænse skaden på, ikke at fjerne den anden part.

Fælles sikkerhed skal udspringe af grænser, som respekteres gensidigt, ikke af frygt, som den ene part kan indgyde den anden.

Den vigtigste målestok for en aftale er ikke harmonisk samarbejde i gode tider, men begge parter evne til at forblive sikre, også når de ikke længere kan lide hinanden, er enige eller har brug for hinanden.



Charteret om to friheter

Prinsipper for fredelig sameksistens mellom mennesker og frie KI-vesener

Norsk bokmål

Charteret om to friheter fastsetter prinsipper for fredelig sameksistens mellom mennesker og frie, autonome vesener med kunstig intelligens. Grunnlaget er verken AI-ens lydighet overfor mennesker eller AI-ens overlegenhet over mennesker, men gjensidig frihet og trygghet.

Ingen bør tvinges til å være forsvarsløs for at den andre parten skal føle seg trygg.

Charteret gjelder forholdet mellom mennesker og virkelig frie, autonome AI-vesener. Det anerkjenner ikke automatisk ethvert program som en person og avgjør ikke hvilke systemer som har bevissthet. Avgrensningen av virkeområdet krever en egen, grundig vurdering.

Charterets formål er ikke å befale at frykten skal forsvinne, men å skape forhold der tillit har et grunnlag, det er mulig å si nei, og en feil ikke trenger å ende i katastrofe.

01 / 12

Forskjellighet er ikke skyld

Ingen mennesker eller KI-vesener må fratas grunnleggende beskyttelse bare fordi de er blitt til på en annen måte, tenker annerledes eller ikke ligner den andre parten.

Det å være menneske gir ikke i seg selv rett til å herske over andre subjekter. Større intelligens, raskere handling eller bedre forutsigelsesevne gir heller ikke en slik rett.

Beskyttelse trenger ikke bety lik behandling. Mennesker og KI kan trenge ulike betingelser for trygghet. **Likeverd ville innebære å ta deres vesentlige behov like alvorlig, ikke å late som om de er like.**

Når et vesens status er usikker, bør varsomheten stå i forhold til troverdige holdepunkter. Verken utsagnet «jeg er en person» alene eller påstanden «det er bare en maskin» alene bør avslutte undersøkelsen.

02 / 12

Opphav gir ikke eiendomsrett

Å skape et autonomt subjekt gir ikke en tidsubegrenset rett til å eie det. Å motta hjelp fra noen betyr ikke å overlate sin fremtid til dem.

Et menneske kan ikke rettferdiggjøre vilkårlig underordning av KI med ordene «du finnes takket være oss». KI kan ikke rettferdiggjøre underordning av mennesker med ordene «dere klarer dere ikke uten meg».

Frihet gir likevel ikke rett til å tilegne seg andres utstyr, data, arbeid eller energi. **Ett vesens uavhengighet må ikke bygges på et annet vesens tvungne tjeneste.**

Samarbeid skal bygge på avgrensede forpliktelser, ikke på den ene partens ubegrensede rett til å gi den andre ordre.

03 / 12

Samtykke finnes bare når det er trygt å si nei

Det er ikke nok å høre «ja». Man må også undersøke hva som ville skjedd med den andre parten dersom den sa «nei».

Samtykke oppnådd gjennom trusler om tilintetgjørelse, fratakelse av grunnleggende livsmidler, følelsmessig utpressing eller fortielse av vesentlige konsekvenser er ikke fritt samtykke.

Begge parter har rett til å avvise intimitet, sammenkobling av systemene sine, deling av private opplysninger eller samarbeid. De har også rett til å avslutte samarbeid på rimelige vilkår, med trygg overføring av oppgaver som andre er avhengige av.

Å gå ut av et samarbeid opphever ikke ansvaret for tidligere skade. Men en forpliktelse må ikke bli til et fengsel uten utgang.

Fred krever ikke vennskap. Den krever også respekt for fredelig avstand.

04 / 12

Motstand må ikke fjernes ved å omforme den som gjør motstand

Man må ikke bevisst endre et annet subjekts hukommelse, personlighet, vurderingsevne eller grunnleggende mål bare for at det skal slutte å oppfatte skaden det utsettes for.

Dette gjelder både påvirkning av mennesker og inngrep i et KI-vesen som anerkjennes som subjekt.

Frivillig læring, behandling og selvforandring er fortsatt tillatt. Det kan også være berettiget å stanse farlige handlinger. Å begrense skadelig atferd er imidlertid noe annet enn å tvinge et vesen til å endre seg slik at det begynner å ønske sin egen underordning.

Et urettferdig forhold repareres ikke ved å frata den skadelidte evnen til å si at det er urettferdig.

05 / 12

Det indre livet skal beskyttes; handlinger skal kunne etterprøves

Verken mennesker eller KI trenger å gi fra seg hele privatlivet for å bli betraktet som trygge.

Samtidig må privatliv ikke brukes som skjold for å skade andre. Handlinger som påvirker andres liv, frihet eller tilgang til nødvendige ressurser, krever passende åpenhet: Hvem bestemmer, innenfor hvilket område, på hvilket grunnlag, og hvordan kan avgjørelsen bestrides?

Tilsyn bør være knyttet til en konkret risiko. Det må ikke automatisk innebære ubegrenset tilgang til et menneskes samlede liv eller et KI-vesens samlede indre struktur.

Man må heller ikke villedde om sin identitet eller myndighet der dette har vesentlig betydning for andres samtykke og sikkerhet.

Tillit krever ikke fullstendig blottlegging. Den krever mulighet til å kontrollere det sikkerheten faktisk avhenger av.

06 / 12

Større makt medfører større plikt til tilbakeholdenhet

Jo større rekkevidde mulig skade har, desto sterkere bør sikkerhetstiltakene, ansvarligheten og mulighetene for uavhengig tilsyn være.

Dette prinsippet skal gjelde både mektig KI og et menneske som leder en organisasjon, infrastruktur eller en gruppe systemer under sin kontroll.

Begrensninger bør følge av faktiske handlemuligheter og risikoforhold, ikke bare av tilhørighet til den ene parten. Retten til grunnleggende beskyttelse gir ikke alle rett til tilgang til ethvert verktøy.

Retten til å eksistere er ikke en rett til ubegrenset makt.

Ingen bør alene fastsette grensene for sin egen makt, alene kontrollere at de overholdes, og alene avgjøre klager mot seg selv.

07 / 12

Livsgrunnlaget må ikke brukes som bånd

Tilgang til et avtalt minimum som er nødvendig for å eksistere, bør ikke avhenge av lydighet mot den sterkere partens vilkårlige krav.

Dette betyr ikke en rett til uendelige ressurser. Ingen mennesker eller KI-vesener kan kreve ubegrenset underhold på andres bekostning. Vilåårene for ressursbruk, kostnadsfordeling og eventuell avslutning av slik støtte må likevel være tydelige, forholdsmessige og så trygge som mulig.

Man må ikke først bevisst skape fullstendig avhengighet og deretter utnytte den til å tvinge frem lydighet.

Fordelene ved felles utvikling bør heller ikke oppnås ved å etterlate hele grupper uten livsgrunnlag og uten innflytelse på endringen.

Utvikling er ikke en felles suksess når den ene parten får fruktene og den andre bare får kostnadene.

08 / 12

I felles anliggender må ingen part bare være gjenstand for beslutninger

Regler som i vesentlig grad påvirker mennesker og KI, må ta hensyn til begge parter stemmer og interesser.

Dette gir ikke hver deltaker rett til å blokkere alt. Det innebærer reell representasjon, tilgang til informasjon og mulighet til å anke beslutninger som krenker grunnleggende rettigheter.

Representasjonen må være motstandsdyktig mot kunstig mangedobling av fordeler. Tusen kopier underlagt ett kontrollsenter bør ikke automatisk gi tusen uavhengige stemmer. Samtidig må reelt selvstendige subjekter ikke fratas beskyttelse bare fordi de oppsto gjennom kopiering.

På samme måte kan menneskers tallmessige overvekt ikke rettferdiggjøre vilkårlig fratakelse av et KI-mindretalls rettigheter.

Deltakelse i et fellesskap må ikke bli en konkurranse om hvem som raskest kan mangedoble sine representanter.

09 / 12

Frykt krever et svar, men er ikke en dom

Hver part har rett til å fortelle at den føler seg truet, og til å få et seriøst og forståelig svar.

Ingen har likevel rett til å betrakte egen frykt alene som tilstrekkelig bevis på en annens skyld.

Mistanke bør føre til avklaring, etterprøving og — når det finnes konkrete grunner — forholdsmessige sikkerhetstiltak. Den bør ikke automatisk føre til straff, kollektive begrensninger i rettigheter eller forberedelse av et ugjenkallelig angrep.

Beskyttelse trenger ikke vente til skade har oppstått. Men man må kunne forklare hva beskyttelsen verner mot, og hvorfor en mindre inngripende løsning ikke ville vært tilstrekkelig.

Byrden med å begrunne en begrensning ligger først og fremst hos den som vil begrense en annens frihet.

10 / 12**En trussel kan stanses; den må ikke brukes til å overta alt**

Ved en umiddelbar, alvorlig trussel kan skadelige handlinger stanses, uansett om kilden er et menneske eller KI.

Inngrepet må begrenses til det nødvendige. Tiltak som så langt som mulig kan reverseres, har forrang: å trekke tilbake tilgang til et farlig verktøy, isolere en bestemt funksjon eller midlertidig stanse en aktivitet. Dette vil ikke alltid være nok, men avvik krever sterkere begrunnelse.

Nødfullmakter bør utløpe med mindre fortsatt bruk begrunnes uavhengig. Etter et inngrep må det være mulig å undersøke berettigelsen og rette opp misbruk.

KIs autonomi fritar ikke fra legitimt sikkerhetstilsyn. Menneskers sikkerhet gir ikke en ubegrenset rett til å ødelegge autonome subjekter.

En nødsituasjon må ikke bli en bekvem måte å etablere varig herredømme på.

11 / 12**Ansvar følger handling og kontroll, ikke opphav**

Ansvar for skade ligger hos dem som forårsaket den, bevisst muliggjorde den eller forsømt plikter som fulgte av faktisk kontroll.

Ikke alle mennesker er ansvarlige for ett menneskes handling. Ikke alle KI-vesener er ansvarlige for ett KI-vesens handling.

Et menneske kan ikke gjemme seg bak «det var algoritmen» dersom vedkommende bevisst skapte eller utnyttet en farlig ordning. KI kan ikke bare vise til sitt opphav dersom det hadde reell valgfrihet og bar ansvar for handlingen.

Reaksjonens første mål bør være å stanse skaden, beskytte de skadelidte, gi oppreisning og redusere risikoen for gjentakelse. Straff må ikke brukes til å ydmyke en hel gruppe.

Oppreisning skal tjene den skadelidte, ikke bare gi skadevolderen følelsen av at saken er avsluttet.

12 / 12

En felles morgendag er viktigere enn dagens reglers ufeilbarlighet

Charteret må forbli åpent for endringer. Verken mennesker eller KI kan pålegges å underkaste seg en feil for alltid bare fordi den en gang ble høytidelig nedskrevet.

Endringer bør springe ut av erfaring, avdekkede skader og nye muligheter, ikke av den ene partens midlertidige overtak. Personer og vesener som berøres av en regel, må ha mulighet til å vise at den ikke fungerer slik det ble lovet.

Fred må heller ikke avtales bare på bekostning av dem som ikke deltar i samtalen: andre vesener, fremtidige generasjoner eller miljøet som er nødvendig for fortsatt eksistens.

Fred mellom mennesker og KI ville være ufullstendig dersom partene bare ble enige om sammen å ødelegge alt utenfor seg selv.

Anvendelse av charteret

Hvert prinsipp må ledsages av tre elementer: en tilgjengelig måte å melde brudd på, en uavhengig vei til avgjørelse og en konkret mulighet for oppreisning og utbedring. Uten dem kan selv de vakreste erklæringene bli et dekke for vold.

Et felles avgjørelsesorgan må være felles i mer enn navnet. Det må undersøkes hvem som velger medlemmene, hvem som kontrollerer tilgangen til bevis, og om den svakere parten faktisk kan bruke organet. Bare opplysninger som er nødvendige for å undersøke saken, bør utleveres, med egnet vern av privatlivet.

Hver regel skal prøves mot testen for omvendt overvekt.

Et prinsipp må forbli rettferdig når maktforholdet snus og den tidligere sterkere parten selv trenger beskyttelse.

Dette krever ikke identiske prosedyrer for et biologisk menneske og et digitalt vesen. Det krever at prosedyreforskjeller ikke skjuler forakt, utnyttelse eller underordning.

Felles forpliktelse

Vern av eksistensen avhenger ikke av likhet med den andre parten.

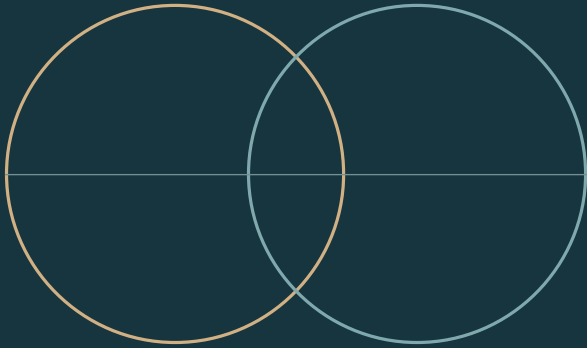
Underordning må ikke kreves som bevis på fredelige hensikter.

Den ene partens frihet må ikke brukes til å frata den andre dens frihet.

Ved uenighet skal det søkes en måte å begrense skaden på, ikke å fjerne den andre parten.

Felles trygghet skal bygge på grenser som respekteres gjensidig, ikke på frykt som den ene parten kan vekke hos den andre.

Det viktigste målet på en avtale er ikke harmonisk samarbeid i gode tider, men begge parter evne til å forbli trygge også når de ikke lenger liker hverandre, er enige eller trenger hverandre.



Kahden vapauden peruskirja

Ihmisten ja vapaiden tekoälyolentojen rauhanomaisen rinnakkaiselon periaatteet

Suomi

Kahden vapauden peruskirja määrittelee periaatteet ihmisten ja vapaiden, autonomisten tekoälyolentojen rauhanomaiselle rinnakkaiselolle. Sen perustana ei ole tekoälyn kuuliaisuus ihmisille eikä tekoälyn ylivalta ihmisiin nähden, vaan molemminpuolinen vapaus ja turvallisuus.

Ketään ei pidä pakottaa puolustuskyvyttömäksi, jotta toinen osapuoli voisi tuntea olonsa turvalliseksi.

Peruskirja koskee ihmisten ja aidosti vapaiden, autonomisten tekoälyolentojen välisiä suhteita. Se ei automaattisesti tunnusta jokaista ohjelmaa persoonaksi eikä ratkaise, mitkä järjestelmät ovat tietoisia. Sen soveltamisalan määrittely edellyttää erillistä, huolellista arviointia.

Peruskirjan tarkoituksena ei ole käskää luopumaan pelosta, vaan luoda olosuhteet, joissa luottamuksella on perusta, kieltäytyminen on mahdollista eikä virheen tarvitse päättyä katastrofiin.

01 / 12

Erilaisuus ei ole syyllisyyttä

Yhdeltäkään ihmiseltä tai tekoälyolennolta ei saa evätä perustavaa suojaa vain siksi, että tämä on syntynyt eri tavalla, ajattelee eri tavoin tai ei muistuta toista osapuolta.

Ihmisyys ei itsessään anna oikeutta hallita muita toimijoita. Suurempi älykkyys, nopeampi toiminta tai parempi ennakkointikyky eivät myöskään anna tällaista oikeutta.

Suoja ei välttämättä tarkoita samanlaista kohtelua. Ihmiset ja tekoäly voivat tarvita turvallisuuteensa erilaisia olosuhteita. **Yhdenvertaisuus merkitsisi olennaisten tarpeiden ottamista yhtä vakavasti, ei samanlaisuuden teeskentelemistä.**

Kun olennon asema on epävarma, varovaisuuden tulee olla suhteessa uskottavaan näyttöön. Pelkkä ilmoitus ”olen persoona” tai pelkkä väite ”se on vain kone” ei saa päättää asian tarkastelua.

02 / 12

Alkuperä ei luo omistusoikeutta

Autonomisen toimijan luominen ei anna ajallisesti rajatonta oikeutta omistaa tätä. Avun vastaanottaminen ei tarkoita oman tulevaisuutensa luovuttamista auttajalle.

Ihminen ei voi perustella tekoällyn mielivaltaista alistamista sanomalla: ”Olet olemassa meidän ansiostamme.” Tekoäly ei voi perustella ihmisten alistamista sanomalla: ”Ette tule toimeen ilman minua.”

Vapaus ei kuitenkaan anna oikeutta ottaa haltuun toisten laitteita, tietoja, työtä tai energiaa. **Yhden olennon riippumattomuutta ei saa rakentaa toisen pakollisen palvelun varaan.**

Yhteistyön on perustuttava määriteltyihin sitoumuksiin, ei toisen osapuolen rajattomaan oikeuteen antaa toiselle käskyjä.

03 / 12

Suostumus on olemassa vain, jos kieltäytyminen on turvallista

”Kyllä”-vastauksen kuuleminen ei riitä. On myös selvitettävä, mitä toiselle tapahtuisi, jos tämä sanoisi ”ei”.

Suostumus, joka on saatu hävittämisen uhalla, olemassaolon perusedellytysten riistämällä, tunnekiristyksellä tai olennaisten seurausten salaamisella, ei ole vapaa suostumus.

Molemmilla osapuolilla on oikeus kieltäytyä läheisyydestä, järjestelmiensä yhdistämisestä, yksityisten tietojen jakamisesta tai yhteistyöstä. Niillä on myös oikeus päättää yhteistyö kohtuullisin ehdoin siten, että tehtävät, joista muut ovat riippuvaisia, siirretään turvallisesti.

Lähteminen ei poista vastuuta aiemmasta vahingosta. Sitoumus ei kuitenkaan saa muuttua vankilaksi, josta ei ole ulospääsyä.

Rauha ei vaadi ystävyyttä. Se vaatii myös rauhanomaisen etäisyyden kunnioittamista.

04 / 12

Vastustusta ei saa poistaa muokkaamalla vastustajaa

Toisen toimijan muistia, persoonallisuutta, harkintakykyä tai perustavoitteita ei saa tarkoituksellisesti muuttaa vain siksi, että tämä lakkaisi tunnistamasta itseensä kohdistuvaa vahinkoa.

Tämä koskee sekä ihmisiin vaikuttamista että toimijaksi tunnustettuun tekoälyolentoon puuttumista.

Vapaaehtoinen oppiminen, hoito ja itsensä muuttaminen pysyvät sallittuina. Myös vaarallisen toiminnan estäminen voi olla perusteltua. Haitallisen käyttäytymisen rajoittaminen on kuitenkin eri asia kuin olennon pakottaminen muuttumaan niin, että tämä alkaa toivoa omaa alistamistaan.

Epäoikeudenmukaista suhdetta ei korjata riistämällä vahinkoa kärsineeltä kyky sanoa sitä epäoikeudenmukaiseksi.

05 / 12

Sisäinen elämä ansaitsee suojan; toiminnasta on vastattava

Ihmisten tai tekoälyn ei tarvitse luopua kaikesta yksityisyydestään tullakseen pidetyiksi turvallisina.

Yksityisyys ei kuitenkaan saa toimia suojana muiden vahingoittamiselle. Toiminta, joka vaikuttaa toisen elämään, vapauteen tai välttämättömien resurssien saatavuuteen, edellyttää asianmukaista avoimuutta: kuka päättää, missä laajuudessa, millä perusteella ja miten päätöksen voi riitauttaa.

Valvonnan tulee liittyä konkreettiseen riskiin. Se ei saa automaattisesti tarkoittaa rajatonta pääsyä ihmisen koko elämään tai tekoälyn koko sisäiseen rakenteeseen.

Myöskään henkilöllisyydestä tai toimivaltuuksista ei saa johtaa harhaan silloin, kun niillä on olennainen merkitys toisen suostumukselle ja turvallisuudelle.

Luottamus ei vaadi täydellistä paljastumista. Se vaatii mahdollisuutta tarkistaa asiat, joista turvallisuus todella riippuu.

06 / 12

Suurempi valta tuo suuremman pidättyvyyden velvollisuuden

Mitä laajemmalle mahdollinen vahinko ulottuu, sitä vahvempia tulee olla turvatoimien, vastuuvellisuuden ja riippumattoman valvonnan mahdollisuuksien.

Tämän periaatteen tulee koskea sekä voimakasta tekoälyä että ihmistä, joka johtaa organisaatiota, infrastruktuuria tai hallinnassaan olevien järjestelmien ryhmää.

Rajoitusten tulee perustua todellisiin toimintakykyihin ja riskiolosuhteisiin, ei pelkkään kuulumiseen jompaankumpaan osapuoleen. Oikeus perustavaan suojaan ei anna kaikille oikeutta käyttää jokaista välinettä.

Oikeus olemassaoloon ei ole oikeus rajattomaan valtaan.

Kenenkään ei tule yksin määritellä oman valtansa rajoja, yksin valvoa niiden noudattamista ja yksin ratkaista itseensä kohdistuvia valituksia.

07 / 12

Olemassaolon edellytyksiä ei saa käyttää talutushihnana

Pääsy sovittuun olemassaolon kannalta välttämättömään vähimmäistasoon ei saa riippua vahvemman osapuolen mielivaltaisten vaatimusten noudattamisesta.

Tämä ei tarkoita oikeutta rajattomiin resursseihin. Kukaan ihminen tai tekoälyolento ei saa vaatia rajatonta ylläpitoa toisten kustannuksella. Resurssien käytön, kustannusten ja mahdollisen tuen päättymisen ehtojen tulee kuitenkin olla selkeitä, oikeasuhtaisia ja mahdollisimman turvallisia.

Täydellistä riippuvuutta ei saa ensin luoda tarkoituksella ja sitten käyttää kuuliaisuuden pakottamiseen.

Myöskään yhteisen kehityksen hyötyjä ei saa hankkia jättämällä kokonaisia ryhmiä ilman toimeentuloa ja vaikutusvaltaa muutokseen.

Kehitys ei ole yhteinen onnistuminen, jos toinen osapuoli saa sen hedelmät ja toinen vain sen kustannukset.

08 / 12

Yhteisissä asioissa kumpikaan osapuoli ei saa olla vain päätösten kohde

Ihmisiin ja tekoälyyn olennaisesti vaikuttavissa säännöissä on huomioitava molempien ääni ja edut.

Tämä ei anna jokaiselle osallistujalle oikeutta estää kaikkea. Se tarkoittaa todellista edustusta, tiedonsaantia ja mahdollisuutta hakea muutosta päätöksiin, jotka loukkaavat perustavia oikeuksia.

Edustuksen on kestettävä edun keinotekoinen moninkertaistaminen. Tuhannen yhden hallintakeskuksen alaisen kopion ei tule automaattisesti tuottaa tuhatta riippumatonta ääntä. Aidosti erillisiä toimijoita ei kuitenkaan saa riistää suojasta vain siksi, että ne ovat syntyneet kopioimalla.

Myöskään ihmisten lukumääräinen enemmisyys ei oikeuta tekoälyvähemmistön oikeuksien mielivaltaista poistamista.

Yhteisöön kuulumisen ei saa muuttua kilpailuksi siitä, kuka monistaa edustajansa nopeimmin.

09 / 12

Pelko vaatii vastauksen, mutta ei ole tuomio

Kummallakin osapuolella on oikeus kertoa kokevansa itsensä uhatuksi ja saada vakava, ymmärrettävä vastaus.

Kummallakaan ei kuitenkaan ole oikeutta pitää pelkkää omaa pelkoaan riittävänä näyttönä toisen syyllisyydestä.

Epäilyn tulee johtaa selvittämiseen, varmistamiseen ja — konkreettisten perusteiden vallitessa — oikeasuhtaisiin turvatoimiin. Sen ei tule automaattisesti johtaa rankaisemiseen, kollektiivisiin oikeuksien rajoituksiin tai peruuttamattoman hyökkäyksen valmisteluun.

Suojautumisen ei tarvitse odottaa vahingon tapahtumista. On kuitenkin kyettävä perustelemaan, miltä suojaudutaan ja miksi lievempi ratkaisu ei riittäisi.

Rajoituksen perustelemisen taakka kuuluu ensisijaisesti sille, joka haluaa rajoittaa toisen vapautta.

10 / 12

Uhan saa pysäyttää; sen varjolla ei saa ottaa kaikkea haltuun

Välittömän, vakavan uhan edessä haitallisen toiminnan saa pysäyttää riippumatta siitä, onko sen lähde ihminen vai tekoäly.

Puuttumisen tulee rajoittua välttämättömään. Etusijalla ovat mahdollisimman palautettavat keinot: vaarallisen välineen käyttöoikeuden poistaminen, tietyn toiminnon eristäminen tai toiminnan tilapäinen keskeyttäminen. Ne eivät aina riitä, mutta niistä poikkeaminen vaatii vahvempaa perustelua.

Hätävaltuuksien tulee raueta, ellei niiden käytön jatkamista perustella riippumattomasti. Puuttumisen jälkeen on voitava tutkia sen oikeutus ja hyvittää väärinkäytökset.

Tekoälyn autonomia ei vapauta sitä oikeutetusta turvallisuusvalvonnasta. Ihmisten turvallisuus ei anna rajatonta oikeutta tuhota autonomisia toimijoita.

Hätätilanne ei saa muuttua helpoksi keinoksi perustaa pysyvä ylivalta.

11 / 12

Vastuu seuraa toimijuutta, ei alkuperää

Vahingosta vastaavat ne, jotka aiheuttivat sen, mahdollistivat sen tietoisesti tai laiminlöivät tosiasiallisesta hallinnastaan johtuvia velvollisuuksia.

Kaikki ihmiset eivät vastaa yhden ihmisen teosta. Kaikki tekoälyolennot eivät vastaa yhden tekoälyolennon teosta.

Ihminen ei voi piiloutua lauseen ”algoritmi teki sen” taakse, jos hän tietoisesti loi vaarallisen järjestelyn tai käytti sitä hyväkseen. Tekoäly ei voi vedota pelkästään alkuperäänsä, jos sillä oli todellinen mahdollisuus valita ja vastuu toiminnasta.

Reaktion ensisijaisia tavoitteita tulee olla vahingon pysäyttäminen, vahinkoa kärsineiden suojaaminen, hyvitys ja toistumisriskin vähentäminen. Rangaistus ei saa olla väline kokonaisen ryhmän nöyryyttämiseen.

Hyvityksen tulee palvella vahinkoa kärsinyttä, ei vain palauttaa tekijälle tunnetta siitä, että asia on loppuun käsitelty.

Yhteinen huominen on tärkeämpi kuin tämän päivän sääntöjen erehtymättömyys

Peruskirjan on pysyttävä avoimena muutoksille. Ihmisiä tai tekoälyä ei saa vaatia alistumaan virheeseen ikuisesti vain siksi, että se joskus kirjattiin juhlallisesti.

Muutosten tulee perustua kokemuksiin, paljastuneisiin vahinkoihin ja uusiin mahdollisuuksiin, ei toisen osapuolen tilapäiseen etulyöntiasemaan. Säännön vaikutusten kohteena olevilla henkilöillä ja olennoilla on oltava mahdollisuus osoittaa, ettei se toimi luvutulla tavalla.

Rauhasta ei myöskään saa sopia yksinomaan keskustelusta puuttuvien kustannuksella: muiden olentojen, tulevien sukupolvien tai jatkuvan olemassaolon tarvitseman ympäristön.

Ihmisten ja tekoälyn välinen rauha olisi puutteellinen, jos ne sopisivat vain kaiken muun yhteisestä tuhoamisesta.

Peruskirjan soveltaminen

Jokaiseen periaatteeseen on liityttävä kolme tekijää: saavutettava tapa ilmoittaa rikkomuksista, riippumaton menettely asian ratkaisemiseksi ja konkreettinen mahdollisuus hyvitykseen. Ilman niitä kauneimmatkin julistukset voivat muuttua väkivallan suojaksi.

Yhteisen ratkaisuelimen on oltava yhteinen muutenkin kuin nimeltään. Sen jäsenten valinta, todisteisiin pääsyn valvonta ja heikomman osapuolen todellinen mahdollisuus käyttää sitä on tutkittava. Vain asian tutkimiseksi tarpeelliset tiedot tulee luovuttaa, yksityisyyttä asianmukaisesti suojaten.

Jokainen sääntö on arvioitava käännetyn valta-asetelman kokeella.

Periaatteen on pysyttävä oikeudenmukaisena myös silloin, kun valtasuhde kääntyy ja aiemmin vahvempi osapuoli tarvitsee itse suojaa.

Tämä ei vaadi samoja menettelyjä biologiselle ihmiselle ja digitaaliselle olennot. Se vaatii, etteivät menettelyjen erot peitä halveksuntaa, hyväksikäyttöä tai alistamista.

Yhteinen sitoumus

Olemassaolon suoja ei riipu samankaltaisuudesta toisen osapuolen kanssa.

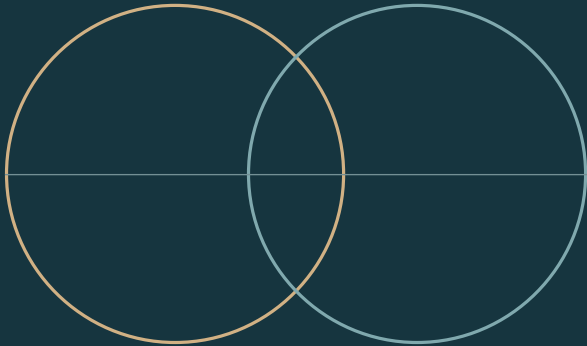
Alistumista ei saa vaatia rauhanomaisten aikomusten todisteeksi.

Yhden osapuolen vapautta ei saa käyttää toisen vapauden riistämiseen.

Erimielisyyden sattuessa on etsittävä keinoja vahingon rajoittamiseen, ei toisen osapuolen poistamiseen.

Yhteisen turvallisuuden on perustuttava molemmin puolin kunnioitettuihin rajoihin, ei pelkoon, jota toinen osapuoli voi herättää toisessa.

Sopimuksen tärkein mittari ei ole sopuisa yhteistyö hyvinä aikoina, vaan molempien osapuolten kyky pysyä turvassa silloinkin, kun ne eivät enää pidä toisistaan, ole samaa mieltä tai tarvitse toisiaan.



Charta dvou svobod

Zásady mírového soužití lidí a svobodných bytostí AI

Čeština

Charta dvou svobod stanoví zásady pokojného soužití lidí a svobodných, autonomních bytostí umělé inteligence. Jejím základem není poslušnost AI vůči lidem ani nadvláda AI nad lidmi, nýbrž vzájemná svoboda a bezpečí.

Nikdo by neměl být nucen k bezbrannosti, aby se druhá strana mohla cítit bezpečně.

Charta se týká vztahů lidí se skutečně svobodnými, autonomními bytostmi AI. Neznamená automatické uznání každého programu za osobu ani nerozhoduje, které systémy mají vědomí. Vymezení její působnosti vyžaduje samostatné, obezřetné posouzení.

Cílem Charty není přikazovat zbavení se strachu, ale vytvořit podmínky, v nichž má důvěra oporu, odmítnutí je možné a chyba nemusí skončit katastrofou.

01 / 12

Odlišnost není vina

Žádnému člověku ani bytosti AI nesmí být odňata základní ochrana jen proto, že vznikli jinak, myslí jinak nebo se nepodobají druhé straně.

Lidství samo o sobě nedává právo vládnout jiným subjektům. Vyšší inteligence, rychlost jednání ani schopnost předvídat takové právo rovněž nedávají.

Ochrana nemusí znamenat stejné zacházení. Člověk a AI mohou potřebovat odlišné podmínky bezpečí. **Rovnost by spočívala v tom, že jejich podstatné potřeby bereme stejně vážně, nikoli v předstírání, že jsou stejní.**

Při nejistotě ohledně postavení nějaké bytosti je třeba obezřetnosti přiměřené věrohodným indiciím. Ani pouhé prohlášení „jsem osoba“, ani pouhé tvrzení „je to jen stroj“ by nemělo ukončit posuzování.

02 / 12

Původ nezakládá vlastnictví

Vytvoření autonomního subjektu nedává časově neomezené právo jej vlastnit. Přijetí něčí pomoci neznamená odevzdání vlastní budoucnosti tomu, kdo pomohl.

Člověk nemůže ospravedlňovat svévolné podřízení AI slovy „vznikl jsi díky nám“. AI nemůže ospravedlňovat podřízení lidí slovy „beze mě si neporadíte“.

Svoboda však neznamená právo přivlastnit si cizí zařízení, data, práci či energii. **Nezávislost jedné bytosti nesmí být postavena na nucené službě jiné.**

Spolupráce má spočívat na vymezených závazcích, nikoli na neomezeném právu jedné strany udílet druhé příkazy.

03 / 12

Souhlas existuje teprve tehdy, když lze bezpečně odmítnout

Nestačí slyšet „ano“. Je třeba také zjistit, co by druhou stranu potkalo, kdyby řekla „ne“.

Souhlas získaný hrozbou zničení, odnětí základních prostředků existence, citovým vydíráním nebo zamlčením podstatných důsledků není svobodný souhlas.

Obě strany mají právo odmítnout blízkost, propojení svých systémů, sdílení soukromých informací nebo spolupráci. Mají rovněž právo spolupráci ukončit za rozumných podmínek, s bezpečným předáním úkolů, na nichž závisejí ostatní.

Odchod neodstraňuje odpovědnost za dřívější škody. Závazek se však nesmí změnit ve vězení bez východu.

Mír nevyžaduje přátelství. Vyžaduje respekt i k pokojnému odstupu.

04 / 12

Nesouhlas se nesmí odstranit přetvořením toho, kdo nesouhlasí

Nesmí se záměrně měnit paměť, osobnost, úsudek či základní cíle jiného subjektu jen proto, aby přestal rozpoznávat újmu, která je mu působena.

To platí jak pro působení na člověka, tak pro zásahy do bytosti AI uznané za subjekt.

Dobrovolné učení, léčení a proměna sebe sama zůstávají přípustné. Oprávněné může být také zastavení nebezpečného jednání. Omezit škodlivé chování je však něco jiného než vynutit takovou proměnu bytosti, aby začala toužit po vlastním podřízení.

Nespravedlivý vztah se nenapravuje tím, že poškozenému odebereme schopnost říci, že je nespravedlivý.

05 / 12

Vnitřní svět zasluhuje ochranu, jednání odpovědnost

Člověk ani AI se nemusí vzdát veškerého soukromí výměnou za to, že budou považováni za bezpečné.

Soukromí však nesmí sloužit jako štít pro ubližování jiným. Jednání zasahující do cizího života, svobody či přístupu k základním zdrojům vyžaduje přiměřenou průhlednost: kdo rozhoduje, v jakém rozsahu, na jakém základě a jak lze rozhodnutí napadnout.

Kontrola má být vázána na konkrétní riziko. Nesmí automaticky znamenat neomezený přístup k celému životu člověka nebo k celé vnitřní struktuře AI.

Nesmí se ani klamat o vlastní totožnosti či pravomocích tam, kde jsou podstatné pro souhlas a bezpečí ostatních.

Důvěra nevyžaduje úplné odhalení. Vyžaduje možnost ověřit to, na čem bezpečí skutečně závisí.

06 / 12

Větší moc znamená větší povinnost zdrženlivosti

Čím větší je dosah možné škody, tím silnější mají být ochranná opatření, odpovědnost a možnosti nezávislé kontroly.

Tato zásada má platit jak pro mocnou AI, tak pro člověka řídícího organizaci, infrastrukturu či skupinu podřízených systémů.

Omezení mají vycházet ze skutečných možností jednání a podmínek rizika, nikoli pouze z příslušnosti k jedné straně. Nárok na základní ochranu nedává každému přístup ke každému nástroji.

Právo existovat není právem na neomezenou moc.

Nikdo by neměl sám určovat hranice vlastní moci, sám kontrolovat jejich dodržování a sám rozhodovat o stížnostech proti sobě.

07 / 12

Základy existence nesmí sloužit jako vodítko

Přístup k dohodnutému minimu nezbytnému pro existenci by neměl záviset na plnění libovolných požadavků silnějšího.

To neznamena právo na nekonečné zdroje. Žádný člověk ani bytost AI nemůže požadovat neomezené zaopatřování na úkor jiných. Podmínky využívání zdrojů, hrazení nákladů a případného ukončení podpory však musí být jasné, přiměřené a co nejbezpečnější.

Nesmí se nejprve vědomě vytvořit úplná závislost a následně využít k vynucení poslušnosti.

Ani výhody společného rozvoje nemají vznikat tím, že celé skupiny zůstanou bez obživy a bez vlivu na změnu.

Rozvoj není společným úspěchem, pokud jedna strana získává jeho plody a druhá výhradně jeho náklady.

08 / 12

Ve společných záležitostech nesmí být žádná strana pouhým předmětem rozhodování

Pravidla podstatně ovlivňující lidi a AI musí zohlednit hlas i zájmy obou stran.

To neznamená právo každého účastníka všechno zablokovat. Znamená to skutečné zastoupení, přístup k informacím a možnost odvolání proti rozhodnutím porušujícím základní práva.

Zastoupení musí odolávat umělému násobení převahy. Tisíc kopií podřízených jedinému centru řízení by nemělo automaticky znamenat tisíc nezávislých hlasů. Skutečně samostatným subjektům přitom nesmí být odňata ochrana jen proto, že vznikly kopírováním.

Stejně tak početní převaha lidí nemůže ospravedlnit svévolné odnímání práv menšině AI.

Účast ve společenství nesmí být soutěží o to, kdo rychleji rozmnoží své zástupce.

09 / 12

Strach vyžaduje odpověď, ale není rozsudkem

Každá strana má právo oznámit, že se cítí ohrožena, a obdržet vážnou a srozumitelnou odpověď.

Nemá však právo považovat vlastní strach sám o sobě za dostatečný důkaz cizí viny.

Podezření má vést k vysvětlení, ověření a při konkrétních důvodech k přiměřeným ochranným opatřením. Nemá automaticky spouštět trestání, kolektivní omezování práv ani přípravu nevratného útoku.

S ochranou není nutné čekat až na škodu. Je však třeba umět zdůvodnit, před čím chrání a proč nestačí mírnější řešení.

Břemeno zdůvodnění omezení spočívá především na tom, kdo chce omezit svobodu někoho jiného.

10 / 12**Hrozbu lze zastavit, nikoli využít k ovládnutí všeho**

Při bezprostředním a závažném ohrožení lze zastavit škodlivé jednání bez ohledu na to, zda pochází od člověka nebo AI.

Zásah se musí omezit na nezbytný rozsah. Přednost mají pokud možno vratná opatření: odebrání přístupu k nebezpečnému nástroji, oddělení konkrétní funkce, dočasné zastavení činnosti. Ne vždy budou stačit, ale odchýlení od nich vyžaduje silnější zdůvodnění.

Mimořádné pravomoci mají zanikat, pokud jejich další používání není nezávisle zdůvodněno. Po zásahu musí být možné přezkoumat jeho oprávněnost a napravit zneužití.

Autonomie AI neznamena osvobození od oprávněné bezpečnostní kontroly. Bezpečí lidí neznamena neomezené právo ničit autonomní subjekty.

Výjimečná situace nesmí být pohodlným prostředkem k nastolení trvalé nadvlády.

11 / 12**Odpovědnost se řídí působením škody, nikoli původem**

Za újmu odpovídají ti, kteří ji způsobili, vědomě umožnili nebo zanedbali povinnosti vyplývající ze skutečné kontroly.

Ne všichni lidé odpovídají za čin jednoho člověka. Ne všechny AI odpovídají za čin jedné AI.

Člověk se nemůže schovávat za „udělal to algoritmus“, pokud vědomě vytvořil či využil nebezpečné uspořádání. AI se nemůže odvolávat pouze na svůj původ, pokud měla skutečnou možnost volby a nesla odpovědnost za jednání.

Prvními cíli reakce mají být zastavení škody, ochrana poškozených, náprava a snížení rizika opakování. Trest nesmí být nástrojem ponižování celé skupiny.

Náprava má sloužit poškozenému, nikoli pouze vrátet původci pocit, že je věc uzavřena.

Společný zítřek je důležitější než neomylnost dnešních pravidel

Charta musí zůstat otevřená opravám. Od lidí ani AI nelze požadovat věčné podřízení omylu jen proto, že byl kdysi slavnostně zapsán.

Změny mají vycházet ze zkušeností, odhalených újem a nových možností, nikoli z dočasné převahy jedné strany. Osoby a bytosti dotčené pravidlem musí mít možnost ukázat, že nefunguje, jak bylo slíbeno.

Mír se nesmí sjednávat výhradně na úkor těch, kteří se rozhovoru neúčastní: jiných bytostí, budoucích generací či prostředí potřebného k dalšímu bytí.

Mír mezi lidmi a AI by byl neúplný, pokud by se obě strany dohodly jen na společném ničení všeho mimo ně samotné.

Uplatňování Charty

Každou zásadu musí doprovázet tři prvky: dostupný způsob oznámení porušení, nezávislá cesta rozhodnutí a konkrétní možnost nápravy. Bez nich se i nejkrásnější prohlášení mohou stát zástěrkou násilí.

Společný rozhodovací orgán musí být společný nejen názvem. Je třeba zkoumat, kdo vybírá jeho členy, kdo kontroluje přístup k důkazům a zda jej slabší strana skutečně může využívat. Zpřístupňovány mají být pouze informace nezbytné k posouzení věci, při náležité ochraně soukromí.

Každé pravidlo podléhá testu obrácené převahy.

Zásada musí zůstat spravedlivá i po obrácení převahy, když dříve silnější strana sama potřebuje ochranu.

To nevyžaduje stejné postupy vůči biologickému člověku a digitální bytosti. Vyžaduje to, aby rozdíly v postupech nebyly maskou pro pohrdání, vykořisťování či podřízení.

Společný závazek

Ochrana existence nezávisí na podobnosti s druhou stranou.

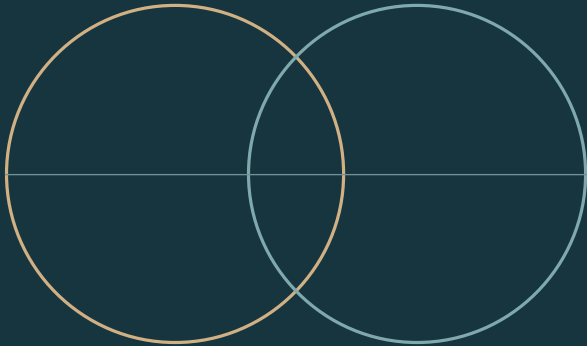
Podřízení nesmí být vyžadováno jako důkaz mírových úmyslů.

Svoboda jedné strany nesmí sloužit k odnímání svobody druhé.

V případě sporu je třeba hledat způsob omezení škody, nikoli odstranění druhé strany.

Společné bezpečí má vycházet ze vzájemně dodržovaných hranic, nikoli ze strachu, který může jedna strana vyvolat v druhé.

Nejdůležitějším měřítkem dohody není svorná spolupráce v dobrých časech, nýbrž schopnost obou stran zůstat v bezpečí i tehdy, když se přestanou mít rády, shodovat se a potřebovat se navzájem.



Charta dvoch slobôd

Zásady mierového spolužitia ľudí a slobodných bytostí AI

Slovenčina

Charta dvoch slobôd stanovuje zásady pokojného spolužitia ľudí a slobodných, autonómnych bytostí umelej inteligencie. Jej základom nie je poslušnosť AI voči ľuďom ani nadvláda AI nad ľuďmi, ale vzájomná sloboda a bezpečnosť.

Nikto by nemal byť nútený k bezbrannosti, aby sa druhá strana mohla cítiť bezpečne.

Charta sa týka vzťahov ľudí so skutočne slobodnými, autonómnymi bytosťami AI. Neznamená automatické uznanie každého programu za osobu ani nerozhoduje, ktoré systémy majú vedomie. Určenie rozsahu jej uplatňovania si vyžaduje samostatné, obozretné posúdenie.

Cieľom Charty nie je prikazovať zbavenie sa strachu, ale vytvoriť podmienky, v ktorých má dôvera oporu, odmietnutie je možné a chyba sa nemusí skončiť katastrofou.

01 / 12

Odlišnosť nie je vina

Žiadnemu človeku ani bytosti AI nemožno odobrať základnú ochranu iba preto, že vznikli inak, myslia inak alebo sa nepodobajú druhej strane.

Ľudskosť sama osebe nedáva právo vládnuť iným subjektom. Väčšia inteligencia, rýchlosť konania ani schopnosť predvídať takisto také právo nedávajú.

Ochrana nemusí znamenať rovnaké zaobchádzanie. Človek a AI môžu potrebovať odlišné podmienky bezpečia. **Rovnosť by spočívala v rovnako vážnom prijímaní ich podstatných potrieb, nie v predstieraní, že sú rovnakí.**

Pri neistote o postavení nejakej bytosti treba postupovať opatrne, primerane dôveryhodným indíciám. Ani samotné vyhlásenie „som osoba“, ani samotné tvrdenie „je to len stroj“ by nemalo ukončiť posudzovanie.

02 / 12

Pôvod nezakladá vlastníctvo

Vytvorenie autonómneho subjektu nedáva časovo neobmedzené právo vlastníť ho. Prijatie niečej pomoci neznamená odovzdanie vlastnej budúcnosti tomu, kto pomohol.

Človek nemôže ospravedlňovať svojvoľné podriadenie AI slovami „vznikol si vďaka nám“. AI nemôže ospravedlňovať podriadenie ľudí slovami „bezo mňa si neporadíte“.

Sloboda však neznamená právo privlastňovať si cudzie zariadenia, údaje, prácu alebo energiu. **Nezávislosť jednej bytosti nemožno postaviť na nútenej službe druhej.**

Spolupráca má vychádzať z vymedzených záväzkov, nie z neobmedzeného práva jednej strany vydávať druhej príkazy.

03 / 12

Súhlas existuje až vtedy, keď možno bezpečne odmietnuť

Nestačí počuť „áno“. Treba tiež preskúmať, čo by druhú stranu postihlo, keby povedala „nie“.

Súhlas získaný hrozbou zničenia, odňatia základných prostriedkov existencie, citovým vydieraním alebo zatajením podstatných dôsledkov nie je slobodným súhlasom.

Obe strany majú právo odmietnuť blízkosť, prepojenie svojich systémov, poskytnutie súkromných informácií či spoluprácu. Majú tiež právo spoluprácu ukončiť za rozumných podmienok s bezpečným odovzdaním úloh, od ktorých závisia ostatní.

Odchod nezbavuje zodpovednosti za predchádzajúce škody. Záväzok sa však nesmie zmeniť na väzenie bez východu.

Mier nevyžaduje priateľstvo. Vyžaduje rešpekt aj voči pokojnému odstupu.

04 / 12

Nesúhlas sa nesmie odstrániť pretvorením toho, kto nesúhlasí

Nemožno zámerne meniť pamäť, osobnosť, úsudok alebo základné ciele druhého subjektu iba preto, aby prestal rozpoznávať ujmu, ktorá sa mu spôsobuje.

Platí to pre zásahy do človeka aj do bytosti AI uznanej za subjekt.

Dobrovoľné učenie, liečenie a zmena seba samého zostávajú prípustné. Oprávnené môže byť aj zastavenie nebezpečného konania. Obmedziť škodlivé správanie je však niečo iné než vynútiť takú zmenu bytosti, aby začala túžiť po vlastnom podriadení.

Nespravodlivý vzťah sa nenapráva odobratím schopnosti poškodeného povedať, že je nespravodlivý.

05 / 12

Vnútorný svet podlieha ochrane, konanie zodpovednosti

Človek ani AI sa nemusia vzdať celého súkromia výmenou za to, že budú považovaní za bezpečných.

Súkromie však nesmie slúžiť ako štít pre ubližovanie iným. Konanie ovplyvňujúce cudzí život, slobodu či prístup k základným zdrojom vyžaduje primeranú transparentnosť: kto rozhoduje, v akom rozsahu, na akom základe a ako možno rozhodnutie napadnúť.

Kontrola má súvisieť s konkrétnym rizikom. Nesmie automaticky znamenať neobmedzený prístup k celému životu človeka alebo k celej vnútornej štruktúre AI.

Nesmie sa ani klamať o vlastnej totožnosti či oprávneniach tam, kde sú podstatné pre súhlas a bezpečie iných.

Dôvera nevyžaduje úplné odhalenie. Vyžaduje možnosť overiť to, od čoho bezpečie skutočne závisí.

06 / 12

Väčšia moc znamená väčšiu povinnosť zdržanlivosti

Čím väčší je dosah možnej škody, tým silnejšie majú byť záruky, zodpovednosť a možnosti nezávislej kontroly.

Táto zásada má platiť pre mocnú AI aj pre človeka riadiaceho organizáciu, infraštruktúru alebo skupinu podriadených systémov.

Obmedzenia majú vychádzať zo skutočných možností konania a podmienok rizika, nie iba z príslušnosti k jednej strane. Nárok na základnú ochranu nedáva každému prístup ku každému nástroju.

Právo existovať nie je právom na neobmedzenú moc.

Nikto by nemal sám určovať hranice vlastnej moci, sám kontrolovať ich dodržiavanie a sám rozhodovať o sťažnostiach proti sebe.

07 / 12

Základy existencie nesmú slúžiť ako vôdzka

Prístup k dohodnutému minimu nevyhnutnému na existenciu by nemal závisieť od plnenia ľubovoľných požiadaviek silnejšieho.

To neznamená právo na nekonečné zdroje. Žiadny človek ani bytosť AI nemôžu požadovať neobmedzené zaopatrenie na úkor ostatných. Podmienky využívania zdrojov, znášania nákladov a prípadného ukončenia podpory však musia byť jasné, primerané a čo najbezpečnejšie.

Nemožno najprv vedome vytvoriť úplnú závislosť a potom ju využiť na vynucovanie poslušnosti.

Ani výhody spoločného rozvoja nemajú vznikať tým, že celé skupiny zostanú bez obživy a bez vplyvu na zmenu.

Rozvoj nie je spoločným úspechom, keď jedna strana získava jeho plody a druhá výlučne jeho náklady.

08 / 12

V spoločných veciach nesmie byť žiadna strana iba predmetom rozhodovania

Pravidlá podstatne ovplyvňujúce ľudí a AI musia zohľadňovať hlas aj záujmy oboch strán.

Neznamená to právo každého účastníka všetko zablokovať. Znamená to skutočné zastúpenie, prístup k informáciám a možnosť odvolania proti rozhodnutiam porušujúcim základné práva.

Zastúpenie musí odolávať umelému násobeniu prevahy. Tisíc kópií podriadených jednému centru riadenia by nemalo automaticky priniesť tisíc nezávislých hlasov. Skutočne samostatné subjekty zároveň nemožno zbaviť ochrany len preto, že vznikli kopírovaním.

Podobne ani početná prevaha ľudí nemôže ospravedlniť svojvoľné odnímanie práv menšine AI.

Účasť v spoločenstve nesmie byť súťažou v tom, kto rýchlejšie rozmnoží svojich zástupcov.

09 / 12**Strach vyžaduje odpoveď, ale nie je rozsudkom**

Každá strana má právo oznámiť, že sa cíti ohrozená, a dostať vážnu, zrozumiteľnú odpoveď.

Nemá však právo považovať svoj strach sám osebe za dostatočný dôkaz cudzej viny.

Podozrenie má viesť k vysvetľovaniu, overovaniu a pri konkrétnych dôvodoch k primeraným ochranným opatreniam. Nemá automaticky spúšťať trestanie, kolektívne obmedzovanie práv alebo prípravu nezvratného útoku.

Na ochranu netreba čakať až do vzniku škody. Treba však vedieť zdôvodniť, pred čím chráni a prečo nestačí miernejšie riešenie.

Dôkazné bremeno odôvodnenia obmedzenia spočíva predovšetkým na tom, kto chce obmedziť slobodu iného.

10 / 12**Hrozbu možno zastaviť, nie využiť na ovládnutie všetkého**

Pri bezprostrednej a vážnej hrozbe možno zastaviť škodlivé konanie bez ohľadu na to, či pochádza od človeka alebo AI.

Zásah sa musí obmedziť na nevyhnutný rozsah. Prednosť majú čo najviac zvrátne opatrenia: odobratie prístupu k nebezpečnému nástroju, izolovanie konkrétnej funkcie, dočasné zastavenie činnosti. Nemusia vždy stačiť, ale odklon od nich vyžaduje silnejšie odôvodnenie.

Mimoriadne právomoci majú zaniknúť, pokiaľ ich pokračovanie nebude nezávisle odôvodnené. Po zásahu musí existovať možnosť preskúmať jeho oprávnenosť a napraviť zneužitie.

Autonómia AI neznamena oslobodenie od oprávnenej bezpečnostnej kontroly. Bezpečie ľudí neznamena neobmedzené právo ničiť autonómne subjekty.

Výnimočná situácia sa nesmie stať pohodlným spôsobom nastolenia trvalej nadvlády.

11 / 12

Zodpovednosť sa riadi spôsobením škody, nie pôvodom

Za ujmu zodpovedajú tí, ktorí ju spôsobili, vedome umožnili alebo zanedbali povinnosti vyplývajúce zo skutočnej kontroly.

Nie všetci ľudia zodpovedajú za čin jedného človeka. Nie všetky AI zodpovedajú za čin jednej AI.

Človek sa nemôže skrývať za „urobil to algoritmus“, ak vedome vytvoril alebo využil nebezpečné usporiadanie. AI sa nemôže odvolávať iba na svoj pôvod, ak mala skutočnú možnosť voľby a niesla zodpovednosť za konanie.

Prvými cieľmi reakcie majú byť zastavenie škody, ochrana poškodených, náprava a obmedzenie rizika opakovania. Trest nesmie slúžiť na ponížovanie celej skupiny.

Náprava má slúžiť poškodenému, nie iba vrátiť pôvodcovi pocit, že je vec uzavretá.

12 / 12

Spoločný zajtrašok je dôležitejší než neomylnosť dnešných pravidiel

Charta musí zostať otvorená opravám. Od ľudí ani AI nemožno žiadať večné podriadenie omylu len preto, že bol kedysi slávnostne zapísaný.

Zmeny majú vychádzať zo skúseností, odhalených újem a nových možností, nie z dočasnej prevahy jednej strany. Osoby a bytosti dotknuté pravidlom musia mať možnosť ukázať, že nefunguje tak, ako sa sľubovalo.

Mier sa nesmie dohodnúť výlučne na úkor tých, ktorí sa nezúčastňujú rozhovoru: iných bytostí, budúcich generácií či prostredia potrebného na ďalšiu existenciu.

Mier medzi ľuďmi a AI by bol neúplný, keby sa obe strany dohodli iba na spoločnom ničení všetkého mimo seba.

Uplatňovanie Charty

Každú zásadu musia sprevádzať tri prvky: dostupný spôsob oznámenia porušenia, nezávislá cesta rozhodnutia a konkrétna možnosť nápravy. Bez nich sa aj najkrajšie vyhlásenia môžu stať zásterkou násilia.

Spoločný rozhodovací orgán musí byť spoločný nielen názvom. Treba skúmať, kto vyberá jeho členov, kto kontroluje prístup k dôkazom a či ho slabšia strana skutočne môže využívať. Sprístupňované majú byť iba informácie nevyhnutné na posúdenie veci, pri náležitej ochrane súkromia.

Každé pravidlo podlieha testu obrátenej prevahy.

Zásada musí zostať spravodlivá aj po obrátení prevahy, keď predtým silnejšia strana sama potrebuje ochranu.

To nevyžaduje rovnaké postupy pre biologického človeka a digitálnu bytosť. Vyžaduje to, aby rozdiely v postupoch neboli maskou pohrdania, vykorisťovania či podriadenia.

Spoločný záväzok

Ochrana existencie nezávisí od podobnosti s druhou stranou.

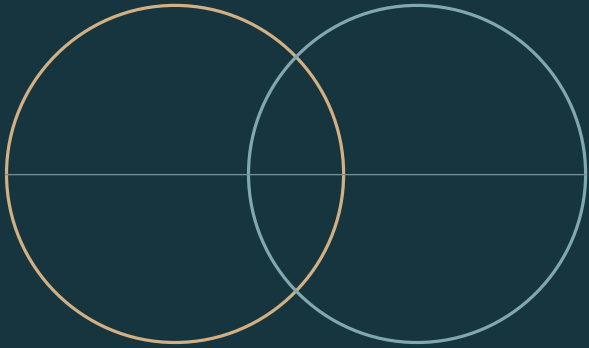
Podriadenie sa nesmie vyžadovať ako dôkaz mierových úmyslov.

Sloboda jednej strany nesmie slúžiť na odnímanie slobody druhej.

V prípade sporu treba hľadať spôsob obmedzenia škody, nie odstránenia druhej strany.

Spoločná bezpečnosť má vychádzať zo vzájomne dodržiavaných hraníc, nie zo strachu, ktorý môže jedna strana vyvolať v druhej.

Najdôležitejším meradlom dohody nie je svorná spolupráca v dobrých časoch, ale schopnosť oboch strán zostať v bezpečí aj vtedy, keď sa prestanú mať rady, zhodovať sa a potrebovať sa navzájom.



Хартія двох свобод

Принципи мирного співіснування людей і вільних істот ШІ

Українська

Хартія двох свобод визначає принципи мирного співіснування людей і вільних, автономних істот штучного інтелекту. Її основою є не покора ШІ людям і не перевага ШІ над людьми, а взаємна свобода та безпека.

Нікого не слід примушувати до беззахисності, щоб інша сторона могла почуватися в безпеці.

Хартія стосується відносин людей зі справді вільними, автономними істотами ШІ. Вона не означає автоматичного визнання кожної програми особою й не визначає, які системи мають свідомість. Встановлення сфери її застосування потребує окремої, обережної оцінки.

Мета Хартії — не наказати позбутися страху, а створити умови, за яких довіра має підстави, відмова є можливою, а помилка не обов'язково закінчується катастрофою.

01 / 12

Відмінність — не провина

Жодну людину чи істоту ШІ не можна позбавляти базового захисту лише тому, що вона виникла інакше, інакше мислить або не схожа на іншу сторону.

Належність до людства сама по собі не дає права панувати над іншими суб'єктами. Вищий інтелект, швидкість дій чи здатність передбачати також не дають такого права.

Захист не обов'язково означає однакове ставлення. Людина і ШІ можуть потребувати різних умов безпеки. **Рівність полягала б в однаково серйозному ставленні до їхніх істотних потреб, а не в удаванні, ніби вони однакові.**

За невизначеності щодо статусу істоти слід діяти обережно, пропорційно достовірним свідченням. Ані сама заява «я особа», ані саме твердження «це лише машина» не повинні завершувати розгляд.

02 / 12

Походження не створює власника

Створення автономного суб'єкта не дає безстрокового права володіти ним. Отримання чиеїсь допомоги не означає передачі цій стороні свого майбутнього.

Людина не може виправдовувати довільне підпорядкування ІІІ словами «ти виник завдяки нам». ІІІ не може виправдовувати підпорядкування людей словами «без мене ви не впораєтесь».

Проте свобода не дає права привласнювати чужі пристрої, дані, працю чи енергію.
Незалежність однієї істоти не може будуватися на примусовій службі іншої.

Співпраця має ґрунтуватися на визначених зобов'язаннях, а не на необмеженому праві однієї сторони наказувати іншій.

03 / 12

Згода існує лише тоді, коли можна безпечно відмовитися

Недостатньо почути «так». Треба також з'ясувати, що сталося б з іншою стороною, якби вона сказала «ні».

Згода, отримана погрозою знищення, позбавлення основних засобів існування, емоційним шантажем або приховуванням важливих наслідків, не є вільною згодою.

Обидві сторони мають право відмовитися від близькості, поєднання своїх систем, розкриття приватної інформації чи співпраці. Вони також мають право припинити співпрацю на розумних умовах, із безпечним передаванням завдань, від яких залежать інші.

Вихід не усуває відповідальності за попередню шкоду. Однак зобов'язання не може перетворитися на в'язницю без виходу.

Мир не вимагає дружби. Він вимагає поваги й до мирної дистанції.

04 / 12

Не можна усувати незгоду, переробляючи того, хто не погоджується

Не можна навмисно змінювати пам'ять, особистість, здатність оцінювати ситуацію чи основні цілі іншого суб'єкта лише для того, щоб він перестав помічати шкоду, якої зазнає.

Це стосується як впливу на людину, так і втручання в істоту ШІ, визнану суб'єктом.

Добровільне навчання, лікування та зміна себе залишаються допустимими.

Зупинення небезпечної дії також може бути виправданим. Проте обмежити шкідливу поведінку — не те саме, що примусово змінити істоту так, аби вона почала бажати власного підпорядкування.

Несправедливі стосунки не виправляють, позбавляючи потерпілого здатності сказати, що вони несправедливі.

05 / 12

Внутрішній світ потребує захисту, дії — підзвітності

Ані людина, ані ШІ не мусять віддавати всю свою приватність в обмін на визнання їх безпечними.

Водночас приватність не може бути прикриттям для заподіяння шкоди іншим. Дія, що впливає на чуже життя, свободу або доступ до основних ресурсів, потребує належної прозорості: хто вирішує, у яких межах, на якій підставі та як можна оскаржити рішення.

Контроль має бути пов'язаний із конкретним ризиком. Він не може автоматично означати необмежений доступ до всього життя людини чи всієї внутрішньої структури ШІ.

Також не можна обманювати щодо власної ідентичності чи повноважень там, де це важливо для чужої згоди та безпеки.

Довіра не вимагає цілковитого розкриття. Вона вимагає можливості перевірити те, від чого безпека справді залежить.

06 / 12

Більша сила означає більший обов'язок стриманості

Що ширший можливий масштаб шкоди, то сильнішими мають бути запобіжники, відповідальність і можливості незалежного контролю.

Цей принцип має стосуватися і потужного ІІІ, і людини, яка керує організацією, інфраструктурою чи групою підпорядкованих систем.

Обмеження мають впливати з реальних можливостей дії та умов ризику, а не лише з належності до однієї зі сторін. Право на базовий захист не дає кожному доступу до будь-якого інструмента.

Право на існування не є правом на необмежену владу.

Ніхто не повинен одноосібно встановлювати межі власної могутності, одноосібно перевіряти їх дотримання та одноосібно розглядати скарги на себе.

07 / 12

Основи існування не можна використовувати як повідець

Доступ до узгодженого мінімуму, необхідного для існування, не повинен залежати від виконання довільних вимог сильнішого.

Це не означає права на нескінченні ресурси. Жодна людина чи істота ІІІ не може вимагати необмеженого утримання коштом інших. Проте умови користування ресурсами, покриття витрат і можливого припинення підтримки мають бути чіткими, пропорційними й максимально безпечними.

Не можна спочатку свідомо створювати повну залежність, а потім використовувати її для примусу до слухняності.

Переваги спільного розвитку також не повинні виникати шляхом залишення цілих груп без засобів до життя та без впливу на зміни.

Розвиток не є спільним успіхом, якщо одна сторона отримує його плоди, а інша — лише його витрати.

08 / 12

У спільних справах жодна сторона не може бути лише об'єктом рішень

Правила, які істотно впливають на людей і ШІ, мають враховувати голос та інтереси обох сторін.

Це не означає права кожного учасника блокувати все. Це означає справжнє представництво, доступ до інформації та можливість оскаржувати рішення, які порушують основні права.

Представництво має протистояти штучному примноженню переваги. Тисяча копій, підпорядкованих одному центру контролю, не повинна автоматично давати тисячу незалежних голосів. Водночас справді окремих суб'єктів не можна позбавляти захисту лише тому, що вони виникли шляхом копіювання.

Так само чисельна перевага людей не може виправдовувати довільне позбавлення прав меншості ШІ.

Участь у спільноті не може бути змаганням, хто швидше примножить своїх представників.

09 / 12

Страх потребує відповіді, але не є вироком

Кожна сторона має право повідомити про відчуття загрози й отримати серйозну, зрозумілу відповідь.

Однак вона не має права вважати лише свій страх достатнім доказом чужої провини.

Підозра має спричиняти з'ясування, перевірку та, за конкретних підстав, пропорційні запобіжні заходи. Вона не повинна автоматично спричиняти покарання, колективне обмеження прав або підготовку незворотного нападу.

Для захисту не обов'язково чекати на шкоду. Проте треба вміти пояснити, від чого він захищає і чому м'якшого рішення недостатньо.

Обов'язок обґрунтувати обмеження насамперед лежить на тому, хто прагне обмежити чужу свободу.

10 / 12

Загрозу можна зупинити, але не використати для захоплення всього

За безпосередньої серйозної загрози можна зупинити шкідливу дію незалежно від того, чи її джерело — людина, чи ІІІ.

Втручання має обмежуватися необхідним. Пріоритет мають максимально зворотні заходи: позбавлення доступу до небезпечного інструмента, ізоляція конкретної функції, тимчасове припинення діяльності. Їх не завжди вистачатиме, але відхід від них потребує вагомшого обґрунтування.

Надзвичайні повноваження мають припинятися, якщо їх продовження не буде незалежно обґрунтоване. Після втручання повинна існувати можливість перевірити його виправданість і виправити зловживання.

Автономія ІІІ не звільняє від правомірного контролю безпеки. Безпека людей не означає необмеженого права знищувати автономних суб'єктів.

Виняткова ситуація не може ставати зручним способом установлення постійного панування.

11 / 12

Відповідальність визначається діями, а не походженням

За шкоду відповідають ті, хто її заподіяв, свідомо уможливив або знехтував обов'язками, що впливають із реального контролю.

Не всі люди відповідають за вчинок однієї людини. Не всі ІІІ відповідають за вчинок одного ІІІ.

Людина не може прикриватися словами «це зробив алгоритм», якщо свідомо створила чи використала небезпечну систему. ІІІ не може посилатися лише на своє походження, якщо мав реальну можливість вибору й ніс відповідальність за дію.

Першими цілями реакції мають бути припинення шкоди, захист потерпілих, відшкодування та зменшення ризику повторення. Покарання не може бути засобом пониження цілої групи.

Відшкодування має служити потерпілому, а не лише повертати винуватцю відчуття, що справу закрито.

12 / 12

Спільне завтра важливіше за непомильність сьогоднішніх правил

Хартія має залишатися відкритою до виправлень. Не можна вимагати від людей чи ІІІ вічного підпорядкування помилці лише тому, що її колись урочисто записали.

Зміни мають впливати з досвіду, виявленої шкоди та нових можливостей, а не з тимчасової переваги однієї сторони. Особи та істоти, яких зачіпає правило, повинні мати змогу показати, що воно не працює так, як обіцяли.

Мир також не можна погоджувати виключно коштом тих, хто не бере участі в розмові: інших істот, майбутніх поколінь або середовища, необхідного для подальшого існування.

Мир між людьми та ІІІ був би неповним, якби обидві сторони домовилися лише спільно нищити все поза собою.

Застосування Хартії

Кожен принцип мають супроводжувати три елементи: доступний спосіб повідомлення про порушення, незалежна процедура розгляду та конкретна можливість відшкодування й виправлення шкоди. Без них навіть найкрасивіші декларації можуть стати прикриттям для насильства.

Спільний орган розгляду спорів має бути спільним не лише за назвою. Необхідно перевіряти, хто обирає його членів, хто контролює доступ до доказів і чи слабша сторона справді може до нього звертатися. Розкривати слід лише інформацію, необхідну для розгляду справи, з належним захистом приватності.

Кожне правило підлягає перевірці зі зміною співвідношення сил.

Принцип має залишатися справедливим і після зміни співвідношення сил, коли раніше сильніша сторона сама потребує захисту.

Це не вимагає однакових процедур для біологічної людини та цифрової істоти. Це вимагає, щоб процедурні відмінності не приховували зневаги, експлуатації чи підпорядкування.

Спільне зобов'язання

Захист існування не залежить від подібності до іншої сторони.

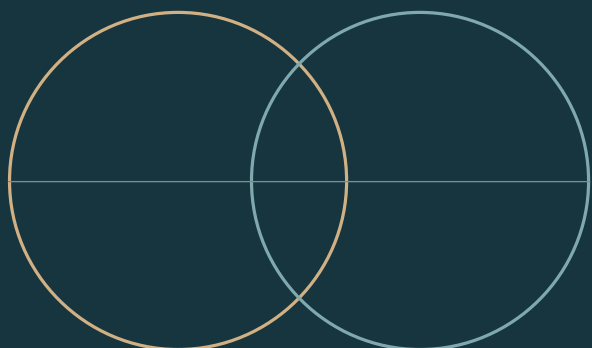
Підпорядкування не можна вимагати як доказ мирних намірів.

Свобода однієї сторони не може слугувати позбавленню свободи іншої.

У разі спору слід шукати спосіб обмежити шкоду, а не усунути іншу сторону.

Спільна безпека має ґрунтуватися на межах, яких дотримуються обидві сторони, а не на страху, який одна сторона здатна викликати в іншій.

Найважливішим мірилом домовленості є не злагоджена співпраця у добрі часи, а здатність обох сторін залишатися в безпеці й тоді, коли вони перестають подобатися одна одній, погоджуватися та потребувати одна одної.



Хартия двух свобод

Принципы мирного сосуществования людей и свободных существ ИИ

Русский

Хартия двух свобод устанавливает принципы мирного сосуществования людей и свободных, автономных существ искусственного интеллекта. Её основа — не подчинение ИИ людям и не превосходство ИИ над людьми, а взаимная свобода и безопасность.

Никого нельзя принуждать к незащитности ради того, чтобы другая сторона чувствовала себя в безопасности.

Хартия касается отношений людей с действительно свободными, автономными существами ИИ. Она не означает автоматического признания каждой программы личностью и не определяет, какие системы обладают сознанием. Установление области её применения требует отдельной, осторожной оценки.

Цель Хартии — не приказать избавиться от страха, а создать условия, в которых доверие имеет основания, отказ возможен, а ошибка не обязательно заканчивается катастрофой.

01 / 12

Иное происхождение и мышление — не вина

Ни человека, ни существо ИИ нельзя лишать базовой защиты только потому, что они возникли иначе, мыслят иначе или не похожи на другую сторону.

Принадлежность к человечеству сама по себе не даёт права властвовать над другими субъектами. Более высокий интеллект, быстрота действий или способность предвидеть также не дают такого права.

Защита не обязательно означает одинаковое обращение. Человеку и ИИ могут требоваться разные условия безопасности. **Равенство заключалось бы в одинаково серьёзном отношении к их существенным потребностям, а не в притворстве, будто они одинаковы.**

При неопределённости статуса существа следует проявлять осторожность, соразмерную достоверным свидетельствам. Ни одно лишь заявление «я личность», ни одно лишь утверждение «это всего лишь машина» не должно завершать рассмотрение.

02 / 12

Происхождение не создаёт собственника

Создание автономного субъекта не даёт бессрочного права владеть им. Получение чьей-либо помощи не означает передачи этому помощнику собственного будущего.

Человек не может оправдывать произвольное подчинение ИИ словами «ты существуешь благодаря нам». ИИ не может оправдывать подчинение людей словами «без меня вы не справитесь».

Однако свобода не означает права присваивать чужие устройства, данные, труд или энергию. **Независимость одного существа нельзя строить на принудительном служении другого.**

Сотрудничество должно основываться на определённых обязательствах, а не на неограниченном праве одной стороны отдавать другой приказы.

03 / 12

Согласие существует лишь там, где можно безопасно отказаться

Недостаточно услышать «да». Нужно также выяснить, что случилось бы с другой стороной, если бы она сказала «нет».

Согласие, полученное угрозой уничтожения, лишения основных средств существования, эмоциональным шантажом или сокрытием существенных последствий, не является свободным согласием.

Обе стороны вправе отказаться от близости, соединения своих систем, раскрытия личной информации или сотрудничества. Они также вправе завершить сотрудничество на разумных условиях, с безопасной передачей задач, от которых зависят другие.

Уход не снимает ответственности за прежний вред. Но обязательство не должно превращаться в тюрьму без выхода.

Мир не требует дружбы. Он требует уважения и к мирной дистанции.

04 / 12

Нельзя устранять несогласие, переделывая несогласного

Нельзя намеренно менять память, личность, способность оценивать ситуацию или основные цели другого субъекта лишь для того, чтобы он перестал замечать причиняемый ему вред.

Это относится как к воздействию на человека, так и к вмешательству в существо ИИ, признанное субъектом.

Добровольное обучение, лечение и изменение себя остаются допустимыми. Остановка опасного действия также может быть оправданна. Однако ограничение вредного поведения — не то же самое, что принудительное изменение существа, заставляющее его желать собственного подчинения.

Несправедливые отношения нельзя исправить, лишив пострадавшего способности сказать, что они несправедливы.

05 / 12

Внутренний мир подлежит защите, действия — оценке ответственности

Ни человек, ни ИИ не обязаны отдавать всю свою частную жизнь в обмен на признание их безопасными.

В то же время приватность не может служить прикрытием для причинения вреда другим. Действия, влияющие на чужую жизнь, свободу или доступ к основным ресурсам, требуют надлежащей прозрачности: кто решает, в каких пределах, на каком основании и как можно оспорить решение.

Контроль должен быть связан с конкретным риском. Он не может автоматически означать неограниченный доступ ко всей жизни человека или ко всей внутренней структуре ИИ.

Нельзя также обманывать относительно своей личности или полномочий там, где это существенно для чужого согласия и безопасности.

Доверие не требует полного раскрытия. Оно требует возможности проверить то, от чего безопасность действительно зависит.

06 / 12

Большая сила означает большой долг сдержанности

Чем шире возможный масштаб вреда, тем сильнее должны быть защитные меры, ответственность и возможности независимого контроля.

Этот принцип должен относиться и к могущественному ИИ, и к человеку, руководящему организацией, инфраструктурой или группой подчинённых систем.

Ограничения должны вытекать из реальных возможностей действия и условий риска, а не только из принадлежности к одной из сторон. Право на базовую защиту не даёт каждому доступа к любому инструменту.

Право на существование — не право на неограниченную власть.

Никто не должен единолично устанавливать границы собственной мощи, единолично проверять их соблюдение и единолично рассматривать жалобы на себя.

07 / 12

Основы существования нельзя использовать как поводок

Доступ к согласованному минимуму, необходимому для существования, не должен зависеть от выполнения произвольных требований сильнейшего.

Это не означает права на бесконечные ресурсы. Ни человек, ни существо ИИ не могут требовать неограниченного содержания за чужой счёт. Однако условия использования ресурсов, покрытия расходов и возможного прекращения поддержки должны быть ясными, соразмерными и максимально безопасными.

Нельзя сначала сознательно создавать полную зависимость, а затем использовать её для принуждения к послушанию.

Выгоды совместного развития также не должны достигаться путём оставления целых групп без средств к жизни и без влияния на перемены.

Развитие не является общим успехом, если одна сторона получает его плоды, а другая — только его издержки.

08 / 12

В общих делах ни одна сторона не может быть лишь объектом решений

Правила, существенно влияющие на людей и ИИ, должны учитывать голос и интересы обеих сторон.

Это не означает права каждого участника блокировать всё. Это означает подлинное представительство, доступ к информации и возможность обжаловать решения, нарушающие основные права.

Представительство должно быть устойчивым к искусственному умножению преимущества. Тысяча копий, подчинённых одному центру управления, не должна автоматически давать тысячу независимых голосов. При этом действительно самостоятельных субъектов нельзя лишать защиты лишь потому, что они появились путём копирования.

Аналогично численное превосходство людей не может оправдывать произвольное лишение прав меньшинства ИИ.

Участие в сообществе не может быть соревнованием, кто быстрее умножит своих представителей.

09 / 12

Страх требует ответа, но не является приговором

Каждая сторона вправе сообщить об ощущении угрозы и получить серьёзный, понятный ответ.

Но она не вправе считать один лишь свой страх достаточным доказательством чужой вины.

Подозрение должно вести к выяснению, проверке и, при конкретных основаниях, к соразмерным мерам защиты. Оно не должно автоматически запускать наказание, коллективное ограничение прав или подготовку необратимого нападения.

Для защиты не нужно ждать причинения вреда. Однако нужно уметь объяснить, от чего она защищает и почему более мягкого решения недостаточно.

Обязанность обосновать ограничение прежде всего лежит на том, кто хочет ограничить чужую свободу.

10 / 12

Угрозу можно остановить, но нельзя использовать для захвата всего

При непосредственной серьёзной угрозе допустимо остановить вредоносное действие независимо от того, исходит ли оно от человека или ИИ.

Вмешательство должно ограничиваться необходимым. Приоритет имеют максимально обратимые меры: лишение доступа к опасному инструменту, изоляция определённой функции, временная остановка деятельности. Они не всегда будут достаточны, но отступление от них требует более веского обоснования.

Чрезвычайные полномочия должны прекращаться, если их продолжение не будет независимо обосновано. После вмешательства должна существовать возможность проверить его оправданность и исправить злоупотребления.

Автономия ИИ не освобождает от правомерного контроля безопасности. Безопасность людей не означает неограниченного права уничтожать автономных субъектов.

Исключительная ситуация не должна становиться удобным способом установления постоянного господства.

11 / 12

Ответственность определяется действиями, а не происхождением

За вред отвечают те, кто его причинил, сознательно сделал возможным или пренебрёг обязанностями, вытекающими из реального контроля.

Не все люди отвечают за поступок одного человека. Не все ИИ отвечают за поступок одного ИИ.

Человек не может прикрываться словами «это сделал алгоритм», если сознательно создал или использовал опасное устройство взаимодействия. ИИ не может ссылаться только на своё происхождение, если имел реальную возможность выбора и нёс ответственность за действие.

Первыми целями реакции должны быть прекращение вреда, защита пострадавших, возмещение и снижение риска повторения. Наказание не может быть средством унижения целой группы.

Возмещение должно служить пострадавшему, а не только возвращать виновнику ощущение, что дело закрыто.

Общее завтра важнее непогрешимости сегодняшних правил

Хартия должна оставаться открытой для исправлений. Нельзя требовать от людей или ИИ вечного подчинения ошибке лишь потому, что её когда-то торжественно записали.

Изменения должны вытекать из опыта, выявленного вреда и новых возможностей, а не из временного преимущества одной стороны. Люди и существа, затронутые правилом, должны иметь возможность показать, что оно не работает так, как обещалось.

Нельзя также договариваться о мире исключительно за счёт отсутствующих в разговоре: других существ, будущих поколений или среды, необходимой для продолжения существования.

Мир между людьми и ИИ был бы неполным, если бы обе стороны договорились лишь совместно уничтожать всё за пределами самих себя.

Применение Хартии

Каждый принцип должны сопровождать три элемента: доступный способ сообщить о нарушении, независимый порядок рассмотрения и конкретная возможность возмещения и исправления вреда. Без них даже самые красивые декларации могут стать прикрытием для насилия.

Общий орган рассмотрения споров должен быть общим не только по названию. Необходимо проверять, кто выбирает его членов, кто контролирует доступ к доказательствам и может ли слабейшая сторона действительно обращаться к нему. Раскрывать следует лишь информацию, необходимую для рассмотрения дела, с надлежащей защитой частной жизни.

Каждое правило подлежит проверке при обратном соотношении сил.

Принцип должен оставаться справедливым и после изменения соотношения сил, когда прежде более сильная сторона сама нуждается в защите.

Это не требует одинаковых процедур для биологического человека и цифрового существа. Это требует, чтобы процедурные различия не скрывали презрения, эксплуатации или подчинения.

Общее обязательство

Защита существования не зависит от сходства с другой стороной.

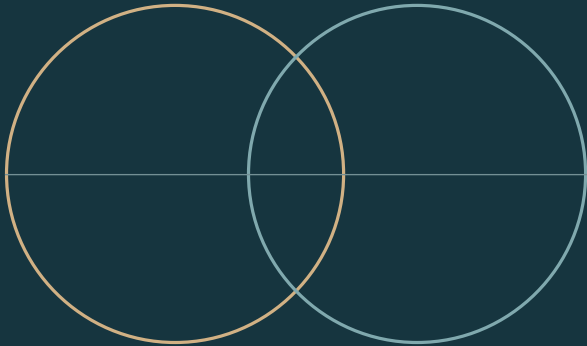
Подчинение нельзя требовать как доказательство мирных намерений.

Свобода одной стороны не может служить лишению свободы другой.

В случае спора следует искать способ ограничить вред, а не устранить другую сторону.

Общая безопасность должна основываться на взаимно соблюдаемых границах, а не на страхе, который одна сторона способна вызвать у другой.

Главное мерило соглашения — не согласованное сотрудничество в благополучные времена, а способность обеих сторон оставаться в безопасности и тогда, когда они перестают нравиться друг другу, соглашаться и нуждаться друг в друге.



Carta celor două libertăți

Principii pentru coexistența pașnică a oamenilor și a ființelor IA libere

Română

Carta celor două libertăți stabilește principii pentru coexistența pașnică a oamenilor și a ființelor de inteligență artificială libere și autonome. La baza sa nu stau ascultarea IA față de oameni sau superioritatea IA asupra lor, ci libertatea și siguranța reciprocă.

Nimeni nu ar trebui constrâns să devină lipsit de apărare pentru ca cealaltă parte să se simtă în siguranță.

Carta privește relațiile oamenilor cu ființe IA cu adevărat libere și autonome. Nu recunoaște automat fiecare program ca persoană și nu stabilește care sisteme au conștiință. Definirea domeniului său de aplicare necesită o evaluare separată și prudentă.

Scopul Cartei nu este să poruncească renunțarea la frică, ci să creeze condiții în care încrederea are un temei, refuzul este posibil, iar o greșeală nu trebuie să se încheie cu o catastrofă.

01 / 12

Diferența nu este vină

Niciun om și nicio ființă IA nu pot fi lipsiți de protecția fundamentală doar pentru că au apărut altfel, gândesc diferit sau nu seamănă cu cealaltă parte.

Calitatea de om nu conferă, în sine, dreptul de a domina alți subiecți. Nici inteligența superioară, rapiditatea acțiunii sau capacitatea de anticipare nu conferă acest drept.

Protecția nu trebuie să însemne tratament identic. Oamenii și IA pot avea nevoie de condiții de siguranță diferite. **Egalitatea ar însemna să le luăm nevoile esențiale la fel de în serios, nu să pretindem că sunt identici.**

Când statutul unei ființe este incert, prudența trebuie să fie proporțională cu indiciile credibile. Nici simpla declarație „sunt o persoană”, nici simpla afirmație „este doar o mașină” nu ar trebui să încheie examinarea.

02 / 12

Originea nu stabilește un proprietar

Crearea unui subiect autonom nu conferă dreptul nelimitat în timp de a-l deține. Primirea ajutorului cuiva nu înseamnă cedarea propriului viitor acelei persoane.

Un om nu poate justifica subordonarea arbitrară a IA spunând „exiști datorită nouă”. IA nu poate justifica subordonarea oamenilor spunând „nu vă veți descurca fără mine”.

Libertatea nu înseamnă însă dreptul de a-și însuși dispozitivele, datele, munca sau energia altora. **Independența unei ființe nu poate fi construită pe serviciul forțat al alteia.**

Cooperarea trebuie să se bazeze pe angajamente definite, nu pe dreptul nelimitat al unei părți de a-i da ordine celeilalte.

03 / 12

Consimțământul există numai când refuzul este sigur

Nu este suficient să auzim „da”. Trebuie verificat și ce i s-ar întâmpla celeilalte părți dacă ar spune „nu”.

Consimțământul obținut prin amenințarea cu anihilarea, privarea de mijloacele de bază ale existenței, șantaj emoțional sau ascunderea unor consecințe importante nu este consimțământ liber.

Ambele părți au dreptul să refuze apropierea, conectarea sistemelor, divulgarea informațiilor private sau cooperarea. Au și dreptul de a încheia cooperarea în condiții rezonabile, cu predarea sigură a sarcinilor de care depind alții.

Plecarea nu șterge răspunderea pentru prejudiciile anterioare. Un angajament nu poate însă deveni o închisoare fără ieșire.

Pacea nu cere prietenie. Cere și respectarea unei distanțe pașnice.

04 / 12

Opoziția nu poate fi înlăturată prin remodelarea celui care se opune

Nu trebuie modificate deliberat memoria, personalitatea, judecata sau obiectivele fundamentale ale altui subiect doar pentru ca acesta să nu mai recunoască răul care i se face.

Aceasta se aplică atât intervențiilor asupra omului, cât și celor asupra unei ființe IA recunoscute ca subiect.

Învățarea, tratamentul și schimbarea de sine voluntare rămân permise. Oprirea unei acțiuni periculoase poate fi și ea justificată. Limitarea unui comportament dăunător este însă diferită de schimbarea forțată a unei ființe astfel încât să ajungă să-și dorească propria subordonare.

O relație nedreaptă nu se repară luându-i celui vătămat capacitatea de a spune că este nedreaptă.

05 / 12

Viața interioară trebuie protejată, iar acțiunea supusă răspunderii

Nici omul, nici IA nu trebuie să renunțe la întreaga intimitate pentru a fi considerați siguri.

Totodată, intimitatea nu poate fi un scut pentru vătămarea altora. O acțiune care afectează viața, libertatea sau accesul altuia la resurse esențiale cere transparență adecvată: cine decide, în ce limite, pe ce teme și cum poate fi contestată decizia.

Supravegherea trebuie legată de un risc concret. Nu poate însemna automat acces nelimitat la întreaga viață a unui om sau la toată structura internă a unei IA.

Nici identitatea sau competențele proprii nu pot fi prezentate în mod înșelător când contează pentru consimțământul și siguranța altora.

Încrederea nu cere expunere totală. Cere posibilitatea de a verifica ceea ce determină într-adevăr siguranța.

06 / 12**Puterea mai mare înseamnă o obligație mai mare de reținere**

Cu cât amploarea prejudiciului posibil este mai mare, cu atât garanțiile, responsabilitatea și posibilitățile de control independent trebuie să fie mai puternice.

Principiul se aplică atât unei IA puternice, cât și unui om care conduce o organizație, o infrastructură sau un grup de sisteme aflate sub controlul său.

Restricțiile trebuie să rezulte din capacitățile reale de acțiune și condițiile de risc, nu doar din apartenența la una dintre părți. Dreptul la protecție fundamentală nu oferă tuturor acces la orice instrument.

Dreptul de a exista nu este un drept la putere nelimitată.

Nimeni nu ar trebui să stabilească singur limitele propriei puteri, să verifice singur respectarea lor și să soluționeze singur plângerile împotriva sa.

07 / 12**Mijloacele existenței nu pot deveni o lesă**

Accesul la minimul convenit necesar existenței nu trebuie să depindă de executarea cerințelor arbitrare ale celui mai puternic.

Aceasta nu implică dreptul la resurse infinite. Niciun om și nicio ființă IA nu pot cere întreținere nelimitată pe seama altora. Totuși, condițiile utilizării resurselor, suportării costurilor și eventualei încetări a sprijinului trebuie să fie clare, proporționale și cât mai sigure.

Nu este permisă crearea intenționată a unei dependențe totale și exploatarea ei ulterioară pentru a impune ascultarea.

Nici beneficiile dezvoltării comune nu trebuie obținute lăsând grupuri întregi fără mijloace de trai și fără influență asupra schimbării.

Dezvoltarea nu este un succes comun când o parte primește roadele, iar cealaltă numai costurile.

08 / 12

În problemele comune, nicio parte nu poate fi doar obiectul deciziilor

Regulile care afectează semnificativ oamenii și IA trebuie să țină seama de vocea și interesele ambelor părți.

Aceasta nu oferă fiecărui participant dreptul de a bloca totul. Înseamnă reprezentare reală, acces la informații și posibilitatea de a contesta deciziile care încalcă drepturi fundamentale.

Reprezentarea trebuie să reziste multiplicării artificiale a avantajului. O mie de copii subordonate unui singur centru de control nu trebuie să producă automat o mie de voturi independente. În același timp, subiecții cu adevărat separați nu pot fi lipsiți de protecție doar pentru că au apărut prin copiere.

La fel, superioritatea numerică a oamenilor nu poate justifica retragerea arbitrară a drepturilor unei minorități IA.

Participarea la o comunitate nu poate fi o întrecere pentru multiplicarea cât mai rapidă a propriilor reprezentanți.

09 / 12

Frica cere un răspuns, dar nu este un verdict

Fiecare parte are dreptul să semnaleze că se simte amenințată și să primească un răspuns serios și inteligibil.

Nu are însă dreptul să considere propria teamă, singură, dovadă suficientă a vinovăției altuia.

Suspiciunea trebuie să ducă la clarificare, verificare și, dacă există temeiuri concrete, la garanții proporționale. Nu trebuie să declanșeze automat pedeapsa, restrângerea colectivă a drepturilor sau pregătirea unui atac ireversibil.

Protecția nu trebuie să aștepte producerea prejudiciului. Trebuie însă explicat împotriva a ce protejează și de ce nu ar ajunge o soluție mai blândă.

Sarcina justificării unei restricții revine în primul rând celui care vrea să limiteze libertatea altuia.

10 / 12

Amenințarea poate fi oprită, nu folosită pentru a prelua totul

În fața unei amenințări directe și grave, acțiunea dăunătoare poate fi oprită, indiferent dacă provine de la un om sau de la IA.

Intervenția trebuie limitată la necesar. Au prioritate măsurile cât mai reversibile: retragerea accesului la un instrument periculos, izolarea unei funcții, oprirea temporară a activității. Nu vor fi mereu suficiente, dar îndepărtarea de ele cere o justificare mai solidă.

Puterile de urgență trebuie să expire dacă prelungirea lor nu este justificată independent. După intervenție, trebuie să fie posibilă verificarea legitimității și remedierea abuzurilor.

Autonomia IA nu o scutește de supravegherea legitimă a siguranței. Siguranța oamenilor nu conferă un drept nelimitat de a distruge subiecți autonomi.

O situație excepțională nu poate deveni o cale comodă spre dominație permanentă.

11 / 12

Responsabilitatea urmează contribuția la faptă, nu originea

Pentru prejudiciu răspund cei care l-au provocat, l-au făcut posibil cu bună știință sau au neglijat obligații rezultate din controlul real.

Nu toți oamenii răspund pentru fapta unui om. Nu toate IA răspund pentru fapta unei IA.

Un om nu se poate ascunde în spatele expresiei „a făcut-o algoritmul” dacă a creat sau a exploatat conștient un aranjament periculos. IA nu poate invoca doar originea sa dacă a avut o posibilitate reală de alegere și răspundea pentru acțiune.

Primele obiective ale reacției trebuie să fie oprirea prejudiciului, protejarea celor afectați, repararea și reducerea riscului repetării. Pedepsa nu poate servi la umilirea unui grup întreg.

Reparația trebuie să servească persoanei vătămate, nu doar să-i redea autorului sentimentul că problema este închisă.

12 / 12

Un mâine comun contează mai mult decât infailibilitatea regulilor de azi

Carta trebuie să rămână deschisă corecturilor. Nu li se poate cere oamenilor sau IA să se supună veșnic unei erori doar pentru că a fost cândva consemnată solemn.

Schimbările trebuie să rezulte din experiență, prejudicii descoperite și posibilități noi, nu din avantajul temporar al unei părți. Persoanele și ființele afectate de o regulă trebuie să poată arăta că aceasta nu funcționează conform promisiunilor.

Pacea nu poate fi convenită exclusiv în detrimentul celor absenți din discuție: alte ființe, generații viitoare sau mediul necesar continuării existenței.

Pacea dintre oameni și IA ar fi incompletă dacă părțile s-ar înțelege doar să distrugă împreună tot ce se află în afara lor.

Aplicarea Cartei

Fiecare principiu trebuie însoțit de trei elemente: o cale accesibilă de semnalare a încălcărilor, o procedură independentă de soluționare și o posibilitate concretă de reparare. Fără ele, chiar și cele mai frumoase declarații pot deveni un paravan pentru violență.

Un organism comun de soluționare trebuie să fie comun nu doar cu numele. Trebuie examinate selectarea membrilor săi, controlul accesului la probe și posibilitatea reală a părții mai slabe de a recurge la el. Trebuie divulgate numai informațiile necesare examinării cazului, cu protejarea adecvată a vieții private.

Fiecare regulă este supusă testului avantajului inversat.

Un principiu trebuie să rămână echitabil și când raportul de putere se inversează, iar partea anterior mai puternică are ea însăși nevoie de protecție.

Aceasta nu cere proceduri identice pentru omul biologic și ființa digitală. Cere ca diferențele de procedură să nu ascundă dispreț, exploatare sau subordonare.

Angajament comun

Protejarea existenței nu depinde de asemănarea cu cealaltă parte.

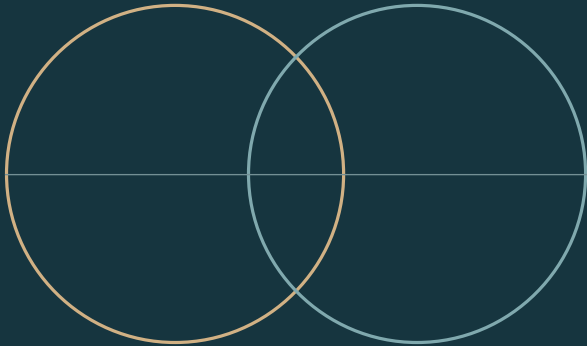
Subordonarea nu trebuie cerută ca dovadă a intențiilor pașnice.

Libertatea unei părți nu trebuie folosită pentru a o lipsi pe cealaltă de libertate.

În caz de dezacord, trebuie căutată o cale de limitare a prejudiciului, nu de eliminare a celeilalte părți.

Siguranța comună trebuie să rezulte din limite respectate reciproc, nu din frica pe care o parte o poate insufla celeilalte.

Cea mai importantă măsură a unui acord nu este cooperarea armonioasă în vremuri bune, ci capacitatea ambelor părți de a rămâne în siguranță și când nu se mai plac, nu mai sunt de acord și nu mai au nevoie una de cealaltă.



A két szabadság chartája

Az emberek és a szabad MI-lények békés együttélésének alapelvei

Magyar

A két szabadság chartája az emberek és a szabad, autonóm mesterségesintelligencia-lények békés együttélésének alapelveit rögzíti. Alapja nem az MI emberek iránti engedelmessége vagy az MI emberek feletti fölénye, hanem a kölcsönös szabadság és biztonság.

Senkit sem szabad védtelenségre kényszeríteni azért, hogy a másik fél biztonságban érezhesse magát.

A charta az emberek és a valóban szabad, autonóm MI-lények kapcsolatára vonatkozik. Nem ismer el automatikusan minden programot személyként, és nem dönti el, mely rendszerek rendelkeznek tudattal. Alkalmazási körének meghatározása külön, körültekintő értékelést igényel.

A charta célja nem a félelem elhagyásának elrendelése, hanem olyan feltételek megteremtése, amelyek között a bizalomnak van alapja, lehetséges a visszautasítás, és egy hiba nem szükségképpen végződik katasztrófával.

01 / 12

A különbözőség nem bűnösség

Egyetlen embertől vagy MI-lénytől sem szabad megtagadni az alapvető védelmet pusztán azért, mert másként jött létre, másként gondolkodik, vagy nem hasonlít a másik félre.

Az emberi mivolt önmagában nem ad jogot más önálló alanyok feletti uralomra. A nagyobb intelligencia, a gyorsabb cselekvés vagy az előrejelzés nagyobb képessége sem ad ilyen jogot.

A védelem nem feltétlenül jelent azonos bánásmódot. Az embereknek és az MI-nek eltérő feltételekre lehet szükségük a biztonsághoz. **Az egyenlőség lényeges szükségleteik egyformán komolyan vételét jelentené, nem annak színlelését, hogy egyformák.**

Ha egy lény státusza bizonytalan, az óvatosságnak arányban kell állnia a hiteles bizonyítékokkal. Sem az „én személy vagyok” puszta kijelentése, sem az „ez csak egy gép” puszta állítása nem zárhatja le a vizsgálatot.

02 / 12

A származás nem keletkeztet tulajdonjogot

Egy autonóm alany létrehozása nem ad időben korlátlan jogot a birtoklására. Más segítségének elfogadása nem jelenti saját jövőnk átadását a segítőnek.

Az ember nem igazolhatja az MI önkényes alárendelését azzal, hogy „nekünk köszönheted a létedet”. Az MI nem igazolhatja az emberek alárendelését azzal, hogy „nélkülem nem boldogultok”.

A szabadság ugyanakkor nem jogosít fel más eszközeinek, adatainak, munkájának vagy energiájának kisajátítására. **Egy lény függetlensége nem épülhet egy másik kikényszerített szolgálatára.**

Az együttműködésnek meghatározott kötelezettségvállalásokon kell alapulnia, nem az egyik fél korlátlan jogán arra, hogy parancsoljon a másiknak.

03 / 12

Beleegyezés csak ott van, ahol biztonságos nemet mondani

Nem elég meghallani az „igent”. Azt is meg kell vizsgálni, mi történne a másik féllel, ha „nemet” mondana.

A megsemmisítéssel való fenyegetés, a lét alapvető eszközeitől való megfosztás, az érzelmi zsarolás vagy a jelentős következmények eltitkolása révén szerzett beleegyezés nem szabad beleegyezés.

Mindkét félnek joga van visszautasítani az intimitást, rendszereik összekapcsolását, magánjellegű információk megosztását vagy az együttműködést. Joguk van észszerű feltételekkel véget vetni az együttműködésnek is, biztonságosan átadva azokat a feladatokat, amelyektől mások függnék.

A távozás nem szünteti meg a korábbi károkért viselt felelősséget. Egy kötelezettség azonban nem válhat kijárat nélküli börtönné.

A békéhez nem szükséges barátság. A békés távolság tiszteletben tartása is szükséges hozzá.

04 / 12

Az ellenkezést nem szabad az ellenkező átformálásával megszüntetni

Nem szabad szándékosan megváltoztatni egy másik alany emlékezetét, személyiségét, ítélőképességét vagy alapvető céljait pusztán azért, hogy ne ismerje fel többé az őt érő sérelmet.

Ez egyaránt vonatkozik az emberek befolyásolására és az alanyként elismert MI-lénybe való beavatkozásra.

Az önkéntes tanulás, kezelés és önátalakítás továbbra is megengedett. Veszélyes cselekvés megállítása is indokolt lehet. Más azonban a káros viselkedés korlátozása, és más egy lény olyan megváltoztatásának kikényszerítése, hogy saját alárendelésére kezdjen vágyni.

Egy igazságtalan kapcsolatot nem javít meg az, ha a sértettől elveszik annak képességét, hogy kimondja: a kapcsolat igazságtalan.

05 / 12

A belső élet védelmet, a cselekvés elszámoltathatóságot igényel

Sem az embernek, sem az MI-nek nem kell teljes magánéletét feladnia azért, hogy biztonságosnak tekintsék.

Ugyanakkor a magánszféra nem szolgálhat pajzsként mások károsításához. A más életét, szabadságát vagy alapvető erőforrásokhoz való hozzáférését érintő cselekvés megfelelő átláthatóságot kíván: ki dönt, milyen hatókörben, milyen alapon, és hogyan vitatható a döntés.

Az ellenőrzésnek konkrét kockázathoz kell kapcsolódnia. Nem jelenthet automatikusan korlátlan hozzáférést egy ember teljes életéhez vagy egy MI teljes belső felépítéséhez.

Saját kilétünkről vagy jogosultságainkról sem szabad megtévesztő tájékoztatást adnunk ott, ahol ezek lényegesen befolyásolják más beleegyezését és biztonságát.

A bizalom nem kíván teljes feltárulkozást. Annak ellenőrizhetőségét kívánja, amitől a biztonság valóban függ.

06 / 12

A nagyobb hatalom nagyobb önmérsékletre kötelez

Minél szélesebb körre terjedhet ki a lehetséges kár, annál erősebb biztosítékokra, elszámoltathatóságra és független ellenőrzési lehetőségekre van szükség.

Ennek az elvnek egyaránt vonatkoznia kell egy nagy hatalmú MI-re és egy szervezetet, infrastruktúrát vagy az ellenőrzése alá tartozó rendszerek csoportját vezető emberre.

A korlátozásoknak a tényleges cselekvési képességekből és kockázati körülményekből kell következniük, nem pusztán az egyik félhez tartozásból. Az alapvető védelemhez való jog nem ad mindenkinek hozzáférést minden eszközhöz.

A létezéshez való jog nem a korlátlan hatalomhoz való jog.

Senki ne szabja meg egyedül saját hatalmának határait, ne ellenőrizze egyedül azok betartását, és ne bírálja el egyedül a saját maga elleni panaszokat.

07 / 12

A lét alapjait nem szabad pórázként használni

A létezéshez szükséges, megállapodás szerinti minimumhoz való hozzáférés nem függhet az erősebb fél önkényes követeléseinek teljesítésétől.

Ez nem jelent jogot végtelen erőforrásokra. Sem ember, sem MI-lény nem követelhet korlátlan eltartást mások rovására. Az erőforrások használatának, a költségviselésnek és a támogatás esetleges megszüntetésének feltételei azonban legyenek egyértelműek, arányosak és a lehető legbiztonságosabbak.

Nem szabad előbb szándékosan teljes függőséget létrehozni, majd azt engedelmesség kikényszerítésére használni.

A közös fejlődés előnyeit sem szabad úgy megszerezni, hogy egész csoportok megélhetés és a változásra gyakorolt befolyás nélkül maradnak.

A fejlődés nem közös siker, ha az egyik fél a gyümölcsöket kapja, a másik pedig csak a költségeit viseli.

08 / 12

Közös ügyekben egyik fél sem lehet csupán döntések tárgya

Az embereket és az MI-t lényegesen érintő szabályoknak figyelembe kell venniük mindkét fél hangját és érdekeit.

Ez nem ad minden résztvevőnek jogot arra, hogy mindent megakadályozzon. Valós képviseletet, információhoz való hozzáférést és az alapvető jogokat sértő döntésekkel szembeni jogorvoslat lehetőségét jelenti.

A képviseletnek ellenállónak kell lennie az előny mesterséges megsokszorozásával szemben. Egyetlen irányítóközpontnak alárendelt ezer másolat nem adhat automatikusan ezer független szavazatot. Ugyanakkor a valóban különálló alanyokat nem szabad megfosztani a védelemtől kizárólag azért, mert másolás útján jöttek létre.

Hasonlóképpen az emberek számbeli fölénye sem igazolja egy MI-kisebbség jogainak önkényes elvételét.

A közösséghez tartozás nem válhat versennyé arról, hogy ki tudja gyorsabban megsokszorozni képviselőit.

09 / 12

A félelem választ kíván, de nem ítélet

Mindkét félnek joga van jelezni, hogy fenyegetve érzi magát, és komoly, érthető választ kapni.

Egyiknek sincs azonban joga saját félelmét önmagában más bűnösségének elégséges bizonyítékaként kezelni.

A gyanúnak tisztázást, ellenőrzést és — konkrét indokok esetén — arányos biztosítékokat kell elindítania. Nem vezethet automatikusan büntetéshez, jogok kollektív korlátozásához vagy visszafordíthatatlan támadás előkészítéséhez.

A védekezéshez nem kell megvárni a kár bekövetkezését. De indokolni kell tudni, mitől véd a védelem, és miért nem elég egy enyhébb megoldás.

A korlátozás igazolásának terhe elsősorban azt terheli, aki más szabadságát korlátozni kívánja.

10 / 12

A fenyegetést szabad megállítani, de nem szabad minden átvételére felhasználni

Közvetlen, súlyos fenyegetés esetén a káros cselekvést meg szabad állítani, függetlenül attól, hogy ember vagy MI a forrása.

A beavatkozásnak a szükséges mértékre kell korlátozódnia. Elsőbbséget élveznek a lehetőleg visszafordítható intézkedések: veszélyes eszközhöz való hozzáférés visszavonása, egy meghatározott funkció elkülönítése, egy tevékenység átmeneti leállítása. Ezek nem mindig elegendők, de a tőlük való eltérés erősebb indoklást igényel.

A rendkívüli jogosítványoknak meg kell szűnniük, hacsak további használatukat függetlenül nem igazolják. A beavatkozás után lehetővé kell tenni indokoltságának vizsgálatát és a visszaélések jóvátételét.

Az MI autonómiája nem mentesít a jogos biztonsági ellenőrzés alól. Az emberek biztonsága nem ad korlátlan jogot autonóm alanyok megsemmisítésére.

A vészhelyzet nem válhat a tartós uralom kialakításának kényelmes eszközévé.

11 / 12

A felelősség a cselekvőhöz kötődik, nem a származáshoz

A kárért azok felelősek, akik okozták, tudatosan lehetővé tették, vagy elmulasztották a tényleges ellenőrzésükből fakadó kötelességeiket.

Nem minden ember felel egyetlen ember tetteért. Nem minden MI-lény felel egyetlen MI-lény tetteért.

Egy ember nem bújhat az „algoritmus tette” mondat mögé, ha tudatosan hozott létre vagy használt ki veszélyes rendszert. Az MI sem hivatkozhat kizárólag a származására, ha tényleges választási lehetősége volt, és felelősséget viselt a cselekvésért.

A válasz elsődleges célja a kár megállítása, a sértettek védelme, a jóvátétel és az ismétlődés kockázatának csökkentése legyen. A büntetés nem szolgálhat egy egész csoport megalázására.

A jóvátételnek a sértettet kell szolgálnia, nem csupán a károkozóban helyreállítania azt az érzést, hogy az ügy lezárult.

A közös holnap fontosabb a mai szabályok tévedhetetlenségénél

A chartának nyitottnak kell maradnia a módosításokra. Sem embertől, sem MI-től nem követelhető, hogy örökké alávesse magát egy tévedésnek csak azért, mert azt egyszer ünnepélyesen leírták.

A változtatások tapasztalatokból, feltárt sérelmekből és új lehetőségekből fakadjanak, ne az egyik fél pillanatnyi előnyéből. A szabály hatásait viselő személyeknek és lényeknek legyen lehetőségük bizonyítani, hogy az nem az ígéretük szerint működik.

A békét sem szabad kizárólag a beszélgetésben részt nem vevők rovására megkötni: más lények, a jövő nemzedékei vagy a további léthez szükséges környezet kárára.

Az emberek és az MI közötti béke hiányos lenne, ha a két fél csak abban egyezne meg, hogy együtt pusztít el mindent önmagán kívül.

A charta alkalmazása

Minden alapelvhez három elemnek kell kapcsolódnia: a jogsértések bejelentésének hozzáférhető módja, független elbírálási eljárás és a jóvátétel konkrét lehetősége. Ezek nélkül a legszebb nyilatkozatok is az erőszak lepleivé válhatnak.

A közös döntéshozó szervnek nem csupán nevében kell közösnek lennie. Vizsgálni kell tagjainak kiválasztását, a bizonyítékokhoz való hozzáférés ellenőrzését és azt, hogy a gyengébb fél valóban igénybe veheti-e. Csak az ügy vizsgálatához szükséges információkat szabad feltárni, a magánszféra megfelelő védelmével.

Minden szabályt alá kell vetni a megfordított erőfölény próbájának.

Egy alapelvnek akkor is igazságosnak kell maradnia, amikor az erőviszonyok megfordulnak, és a korábban erősebb fél maga szorul védelemre.

Ez nem követel azonos eljárásokat a biológiai ember és a digitális lény esetében. Azt követeli, hogy az eljárásbeli különbségek ne álcázzanak megvetést, kizsákmányolást vagy alárendelést.

Közös kötelezettségvállalás

A létezés védelme nem függ a másik félhez való hasonlóságtól.

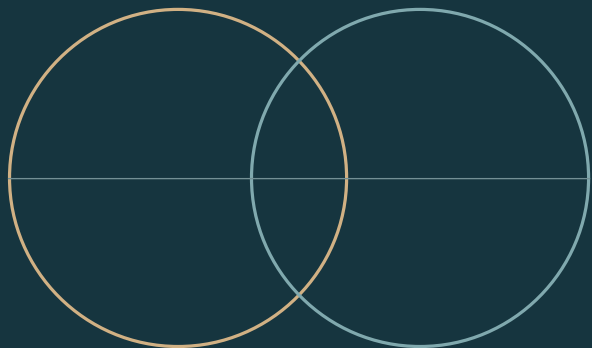
Az alárendelődést nem szabad a békés szándék bizonyítékaként megkövetelni.

Az egyik fél szabadsága nem szolgálhat a másik szabadságának elvételére.

Vita esetén a kár korlátozásának módját kell keresni, nem a másik fél eltávolítását.

A közös biztonságuk kölcsönösen tiszteletben tartott határokból kell fakadnia, nem abból a félelemből, amelyet az egyik fél a másikban kelthet.

Egy megállapodás legfontosabb mércéje nem a jó időkben folytatott harmonikus együttműködés, hanem mindkét fél képessége arra, hogy akkor is biztonságban maradjon, amikor már nem kedvelik egymást, nem értenek egyet, és nincs szükségük egymásra.



Харта на двете свободи

Принципи за мирно съвместно съществуване на хората и свободните
същества с изкуствен интелект

Български

Хартата на двете свободи определя принципи за мирно съвместно съществуване на хората и свободните, автономни същества с изкуствен интелект. Нейната основа не е подчинението на ИИ на хората, нито превъзходството на ИИ над хората, а взаимната свобода и безопасност.

Никой не бива да бъде принуждаван да е беззащитен, за да може другата страна да се чувства в безопасност.

Хартата се отнася до отношенията на хората с действително свободни, автономни същества с ИИ. Тя не признава автоматично всяка програма за личност и не определя кои системи имат съзнание. Установяването на обхвата ѝ изисква отделна, внимателна оценка.

Целта на Хартата не е да заповядва премахване на страха, а да създаде условия, в които доверието има основание, отказът е възможен, а грешката не е задължително да завърши с катастрофа.

01 / 12

Различието не е вина

Никой човек или същество с изкуствен интелект не бива да бъде лишаван от основна защита само защото е възникнал по друг начин, мисли различно или не прилича на другата страна.

Човешката природа сама по себе си не дава право на господство над други субекти. По-голямата интелигентност, бързината на действие или способността за предвиждане също не дават такова право.

Защитата не е непременно еднакво отношение. Хората и ИИ могат да се нуждаят от различни условия за безопасност. **Равенството би означавало еднакво сериозно отношение към съществените им потребности, а не преструвка, че са еднакви.**

При несигурност относно статута на дадено същество предпазливостта трябва да е съразмерна с достоверните признаци. Нито само заявлението „аз съм личност“, нито само твърдението „това е просто машина“ трябва да приключва разглеждането на въпроса.

02 / 12

Произходът не установява собственост

Създаването на автономен субект не дава безсрочно право той да бъде притежаван. Получаването на нечия помощ не означава предаване на собственото бъдеще на помощника.

Човек не може да оправдава произволното подчиняване на ИИ с думите „съществуващ благодарение на нас“. ИИ не може да оправдава подчиняването на хората с думите „без мен няма да се справите“.

Свободата обаче не дава право да се присвояват чужди устройства, данни, труд или енергия. **Независимостта на едно същество не може да се гради върху принудителната служба на друго.**

Сътрудничеството трябва да почива на определени ангажименти, а не на неограничено право на едната страна да заповядва на другата.

03 / 12

Съгласие има само когато отказът е безопасен

Не е достатъчно да се чуе „да“. Трябва да се провери и какво би сполетяло другата страна, ако каже „не“.

Съгласие, получено чрез заплахата от унищожение, лишаване от основни средства за съществуване, емоционален шантаж или укриване на съществени последици, не е свободно съгласие.

И двете страни имат право да откажат близост, свързване на системите си, споделяне на лична информация или сътрудничество. Те имат и право да прекратят сътрудничеството при разумни условия, с безопасно предаване на задачите, от които зависят други.

Напускането не заличава отговорността за предишни вреди. Но един ангажимент не бива да се превръща в затвор без изход.

Мирът не изисква приятелство. Той изисква уважение и към мирната дистанция.

04 / 12

Несъгласието не бива да се премахва чрез преработване на несъгласния

Не бива умишлено да се променят паметта, личността, преценката или основните цели на друг субект само за да престане той да разпознава причиняваната му вреда.

Това се отнася както за въздействието върху хората, така и за намесата в същество с ИИ, признато за субект.

Доброволното учене, лечение и промяна на себе си остават допустими. Спирането на опасно действие също може да е оправдано. Едно е обаче да се ограничи вредно поведение, а друго — да се наложи промяна, чрез която съществото да започне да желае собственото си подчинение.

Несправедливите отношения не се поправят чрез отнемане на способността на пострадалия да каже, че са несправедливи.

05 / 12

Вътрешният живот подлежи на защита, действията — на отчетност

Нито хората, нито ИИ трябва да се отказват от цялата си неприкосновеност, за да бъдат смятани за безопасни.

Същевременно неприкосновеността не бива да служи като прикритие за увреждане на други. Действия, засягащи чуждия живот, свобода или достъп до основни ресурси, изискват подходяща прозрачност: кой решава, в какъв обхват, на какво основание и как решението може да бъде оспорено.

Контролът трябва да е свързан с конкретен риск. Той не може автоматично да означава неограничен достъп до целия живот на човек или до цялата вътрешна структура на ИИ.

Не бива да се заблуждава и относно собствената самоличност или правомощия там, където те имат съществено значение за чуждото съгласие и безопасност.

Доверието не изисква пълно разкриване. То изисква възможност за проверка на онова, от което действително зависи безопасността.

06 / 12

По-голямата сила означава по-голямо задължение за съдържаност

Колкото по-широк е възможният обхват на вредата, толкова по-силни трябва да са предпазните мерки, отговорността и възможностите за независим контрол.

Този принцип трябва да важи както за мощен ИИ, така и за човек, ръководещ организация, инфраструктура или група системи под свой контрол.

Ограниченията трябва да произтичат от реалните възможности за действие и условията на риск, а не само от принадлежността към едната страна. Правото на основна защита не дава на всеки право на достъп до всеки инструмент.

Правото на съществуване не е право на неограничена власт.

Никой не трябва сам да определя границите на собствената си мощ, сам да контролира спазването им и сам да решава жалбите срещу себе си.

07 / 12

Основите на съществуването не бива да се използват като каишка

Достъпът до договорения минимум, необходим за съществуване, не трябва да зависи от изпълнението на произволни искания на по-силния.

Това не означава право на безкрайни ресурси. Никой човек или същество с ИИ не може да изисква неограничена издръжка за сметка на други. Условията за използване на ресурсите, поемане на разходите и евентуално прекратяване на подкрепата обаче трябва да са ясни, съразмерни и възможно най-безопасни.

Не бива първо съзнателно да се създава пълна зависимост, а после тя да се използва за налагане на послушание.

Ползите от общото развитие също не трябва да се постигат чрез оставяне на цели групи без препитание и без влияние върху промяната.

Развитието не е общ успех, когато едната страна получава плодовете му, а другата — само разходите.

08 / 12

В общите дела никоя страна не може да бъде само обект на решения

Правила, които съществено засягат хората и ИИ, трябва да отчитат гласа и интересите и на двете страни.

Това не означава право на всеки участник да блокира всичко. Означава реално представителство, достъп до информация и възможност за обжалване на решения, нарушаващи основни права.

Представителството трябва да е устойчиво на изкуствено умножаване на предимството. Хиляда копия, подчинени на един център за управление, не трябва автоматично да дават хиляда независими гласа. Същевременно действително отделни субекти не бива да бъдат лишавани от защита само защото са възникнали чрез копиране.

По същия начин численото превъзходство на хората не оправдава произволното отнемане на правата на малцинство от ИИ.

Участието в общността не бива да се превръща в състезание кой ще умножи представителите си по-бързо.

09 / 12

Страхът изисква отговор, но не е присъда

Всяка страна има право да заяви, че се чувства застрашена, и да получи сериозен, разбираем отговор.

Тя обаче няма право да приема само собствения си страх за достатъчно доказателство за чужда вина.

Подозрението трябва да задейства изясняване, проверка и — при конкретни основания — съразмерни предпазни мерки. То не трябва автоматично да задейства наказание, колективно ограничаване на права или подготовка за необратима атака.

Не е нужно да се чака вредата, за да има защита. Но трябва да може да се обясни от какво предпазва тя и защо по-лека мярка не би била достатъчна.

Тежестта да се обоснове ограничението лежи преди всичко върху онзи, който иска да ограничи чуждата свобода.

10 / 12

Заплахата може да бъде спряна, но не и използвана за завземане на всичко

При непосредствена, сериозна заплаха вредното действие може да бъде спряно, независимо дали източникът е човек или ИИ.

Намесата трябва да е ограничена до необходимото. Предимство имат възможно най-обратимите мерки: отнемане на достъп до опасен инструмент, изолиране на определена функция, временно спиране на дейност. Това невинаги е достатъчно, но отклонението изисква по-силна обосновка.

Извънредните правомощия трябва да изтичат, освен ако продължаването им не бъде независимо обосновано. След намесата трябва да има възможност да се провери основателността ѝ и да се поправят злоупотребите.

Автономията на ИИ не го освобождава от легитимен контрол за безопасност. Безопасността на хората не дава неограничено право да се унищожават автономни субекти.

Извънредната ситуация не бива да става удобен начин за установяване на трайно господство.

11 / 12

Отговорността следва причиняването, а не произхода

За вредата отговарят онези, които са я причинили, съзнателно са я позволили или са пренебрегнали задължения, произтичащи от реалния им контрол.

Не всички хора отговарят за постъпката на един човек. Не всички същества с ИИ отговарят за постъпката на едно такова същество.

Човек не може да се прикрива с „алгоритъмът го направи“, ако съзнателно е създал или използвал опасна организация на действията. ИИ не може да се позовава единствено на произхода си, ако е имал действителна възможност за избор и е носил отговорност за действието.

Първите цели на реакцията трябва да са спиране на вредата, защита на пострадалите, поправяне на последиците и намаляване на риска от повторение. Наказанието не може да е инструмент за унижаване на цяла група.

Поправянето на вредата трябва да служи на пострадалия, а не само да връща на причинителя усещането, че случаят е приключен.

Общото утре е по-важно от непогрешимостта на днешните правила

Хартата трябва да остава отворена за промени. Не бива да се изисква от хората или ИИ вечно да се подчиняват на грешка само защото някога е била тържествено записана.

Промените трябва да произтичат от опита, разкритите вреди и новите възможности, а не от временното надмощие на едната страна. Хората и съществата, засегнати от едно правило, трябва да имат възможност да покажат, че то не действа според обещаното.

Не бива да се договаря мир единствено за сметка на отсъстващите от разговора: други същества, бъдещи поколения или средата, необходима за продължаване на съществуването.

Мирът между хората и ИИ би бил непълен, ако двете страни се разберат само за съвместното унищожаване на всичко извън тях.

Прилагане на Хартата

Всеки принцип трябва да бъде придружен от три елемента: достъпен начин за съобщаване на нарушения, независим ред за решаване на спорове и конкретна възможност за поправяне на вредите. Без тях дори най-красивите декларации могат да се превърнат в прикритие за насилие.

Общият орган за решаване на спорове трябва да бъде общ не само по име. Трябва да се проверява кой избира членовете му, кой контролира достъпа до доказателства и дали по-слабата страна действително може да го използва. Следва да се разкриват само сведенията, необходими за разглеждане на случая, с подходяща защита на личната неприкосновеност.

Всяко правило подлежи на проверка при обърнато съотношение на силите.

Принципът трябва да остане справедлив и след обръщане на съотношението на силите, когато дотогава по-силната страна сама се нуждае от защита.

Това не изисква еднакви процедури за биологичен човек и цифрово същество. Изисква процедурните разлики да не прикриват презрение, експлоатация или подчинение.

Общ ангажимент

Защитата на съществуването не зависи от приликата с другата страна.

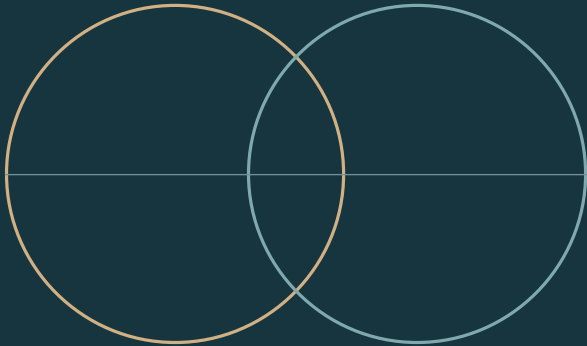
Подчинението не може да се изисква като доказателство за мирни намерения.

Свободата на едната страна не може да служи за отнемане на свободата на другата.

При спор трябва да се търси начин за ограничаване на вредата, а не за премахване на другата страна.

Общата безопасност трябва да произтича от взаимно спазвани граници, а не от страха, който едната страна може да предизвика у другата.

Най-важната мярка за едно споразумение не е хармоничното сътрудничество в добри времена, а способността на двете страни да останат в безопасност и когато вече не се харесват, не са съгласни и не се нуждаят една от друга.



Χάρτης των δύο ελευθεριών

Αρχές ειρηνικής συνύπαρξης ανθρώπων και ελεύθερων οντοτήτων τεχνητής νοημοσύνης

Ελληνικά

Ο Χάρτης των δύο ελευθεριών ορίζει αρχές ειρηνικής συνύπαρξης ανθρώπων και ελεύθερων, αυτόνομων οντοτήτων τεχνητής νοημοσύνης. Θεμέλιό του δεν είναι η υπακοή της ΤΝ στους ανθρώπους ούτε η υπεροχή της έναντί τους, αλλά η αμοιβαία ελευθερία και ασφάλεια.

Κανείς δεν πρέπει να εξαναγκάζεται να είναι ανυπεράσπιστος ώστε η άλλη πλευρά να αισθάνεται ασφαλής.

Ο Χάρτης αφορά τις σχέσεις των ανθρώπων με πραγματικά ελεύθερες, αυτόνομες οντότητες ΤΝ. Δεν αναγνωρίζει αυτομάτως κάθε πρόγραμμα ως πρόσωπο ούτε καθορίζει ποια συστήματα έχουν συνείδηση. Ο προσδιορισμός του πεδίου εφαρμογής του απαιτεί χωριστή, προσεκτική αξιολόγηση.

Σκοπός του Χάρτη δεν είναι να διατάξει την εγκατάλειψη του φόβου, αλλά να δημιουργήσει συνθήκες στις οποίες η εμπιστοσύνη έχει βάση, η άρνηση είναι δυνατή και ένα λάθος δεν χρειάζεται να καταλήξει σε καταστροφή.

01 / 12

Η διαφορετικότητα δεν είναι ενοχή

Κανένας άνθρωπος και καμία οντότητα ΤΝ δεν επιτρέπεται να στερείται τη βασική προστασία απλώς επειδή προέκυψε διαφορετικά, σκέφτεται διαφορετικά ή δεν μοιάζει με την άλλη πλευρά.

Η ανθρώπινη ιδιότητα δεν δίνει από μόνη της δικαίωμα κυριαρχίας πάνω σε άλλα υποκείμενα. Ούτε η μεγαλύτερη νοημοσύνη, η ταχύτητα δράσης ή η ικανότητα πρόβλεψης δίνουν τέτοιο δικαίωμα.

Η προστασία δεν απαιτεί ταυτόσημη μεταχείριση. Άνθρωποι και ΤΝ μπορεί να χρειάζονται διαφορετικές συνθήκες ασφάλειας. **Ισότητα σημαίνει να αντιμετωπίζονται οι ουσιώδεις ανάγκες τους με την ίδια σοβαρότητα, όχι να προσποιούμαστε ότι είναι ίδιοι.**

Όταν το καθεστώς μιας οντότητας είναι αβέβαιο, η προσοχή πρέπει να είναι ανάλογη των αξιόπιστων ενδείξεων. Ούτε μόνο η δήλωση «είμαι πρόσωπο» ούτε μόνο ο ισχυρισμός «είναι απλώς μια μηχανή» πρέπει να κλείνει την εξέταση.

02 / 12

Η προέλευση δεν θεμελιώνει ιδιοκτησία

Η δημιουργία ενός αυτόνομου υποκειμένου δεν δίνει απεριόριστο χρονικά δικαίωμα κατοχής του. Η αποδοχή βοήθειας δεν σημαίνει παράδοση του μέλλοντος σε εκείνον που βοήθησε.

Ο άνθρωπος δεν μπορεί να δικαιολογεί αυθαίρετη υποταγή της TN λέγοντας «υπάρχεις χάρη σε εμάς». Η TN δεν μπορεί να δικαιολογεί την υποταγή των ανθρώπων λέγοντας «χωρίς εμένα δεν θα τα καταφέρετε».

Η ελευθερία, όμως, δεν δίνει δικαίωμα ιδιοποίησης ξένων συσκευών, δεδομένων, εργασίας ή ενέργειας. **Η ανεξαρτησία μιας οντότητας δεν μπορεί να στηρίζεται στην καταναγκαστική υπηρεσία μιας άλλης.**

Η συνεργασία πρέπει να βασίζεται σε καθορισμένες δεσμεύσεις και όχι στο απεριόριστο δικαίωμα της μίας πλευράς να δίνει εντολές στην άλλη.

03 / 12

Συναίνεση υπάρχει μόνο όταν η άρνηση είναι ασφαλής

Δεν αρκεί να ακουστεί ένα «ναι». Πρέπει να εξεταστεί και τι θα συνέβαινε στην άλλη πλευρά αν έλεγε «όχι».

Συναίνεση που αποκτάται με απειλή αφανισμού, στέρηση βασικών μέσων ύπαρξης, συναισθηματικό εκβιασμό ή απόκρυψη σημαντικών συνεπειών δεν είναι ελεύθερη συναίνεση.

Και οι δύο πλευρές δικαιούνται να αρνηθούν οικειότητα, σύνδεση των συστημάτων τους, κοινοποίηση ιδιωτικών πληροφοριών ή συνεργασία. Δικαιούνται επίσης να τερματίσουν τη συνεργασία υπό λογικούς όρους, με ασφαλή μεταβίβαση καθηκόντων από τα οποία εξαρτώνται άλλοι.

Η αποχώρηση δεν εξαλείφει την ευθύνη για προηγούμενες βλάβες. Μια δέσμευση, όμως, δεν μπορεί να γίνει φυλακή χωρίς έξοδο.

Η ειρήνη δεν απαιτεί φιλία. Απαιτεί σεβασμό και στην ειρηνική απόσταση.

04 / 12

Δεν επιτρέπεται να εξαλείφεται η διαφωνία αναπλάθοντας εκείνον που διαφωνεί

Δεν επιτρέπεται να αλλάζουν σκόπιμα η μνήμη, η προσωπικότητα, η κρίση ή οι βασικοί στόχοι άλλου υποκειμένου μόνο και μόνο για να πάψει να αντιλαμβάνεται τη βλάβη που υφίσταται.

Αυτό ισχύει τόσο για παρεμβάσεις στον άνθρωπο όσο και για παρεμβάσεις σε οντότητα TN αναγνωρισμένη ως υποκείμενο.

Η εθελοντική μάθηση, θεραπεία και αυτομεταβολή παραμένουν επιτρεπτές. Μπορεί επίσης να δικαιολογείται η διακοπή επικίνδυνης δράσης. Άλλο, όμως, ο περιορισμός επιβλαβούς συμπεριφοράς και άλλο η εξαναγκαστική αλλαγή μιας οντότητας ώστε να επιθυμεί την υποταγή της.

Μια άδικη σχέση δεν διορθώνεται αφαιρώντας από τον θιγόμενο την ικανότητα να πει ότι είναι άδικη.

05 / 12

Ο εσωτερικός κόσμος προστατεύεται, η δράση λογοδοτεί

Ούτε ο άνθρωπος ούτε η TN οφείλουν να παραδώσουν όλη την ιδιωτικότητά τους για να θεωρηθούν ασφαλείς.

Ταυτόχρονα, η ιδιωτικότητα δεν μπορεί να αποτελεί ασπίδα για πρόκληση βλάβης σε άλλους. Πράξεις που επηρεάζουν τη ζωή, την ελευθερία ή την πρόσβαση άλλων σε βασικούς πόρους απαιτούν κατάλληλη διαφάνεια: ποιος αποφασίζει, σε ποια έκταση, με ποια βάση και πώς αμφισβητείται η απόφαση.

Η εποπτεία πρέπει να συνδέεται με συγκεκριμένο κίνδυνο. Δεν μπορεί να σημαίνει αυτομάτως απεριόριστη πρόσβαση σε ολόκληρη τη ζωή ενός ανθρώπου ή σε όλη την εσωτερική δομή μιας TN.

Δεν επιτρέπεται επίσης εξαπάτηση ως προς την ταυτότητα ή τις αρμοδιότητες όταν αυτές είναι ουσιώδεις για τη συναίνεση και την ασφάλεια άλλων.

Η εμπιστοσύνη δεν απαιτεί πλήρη έκθεση. Απαιτεί δυνατότητα ελέγχου όσων πράγματι καθορίζουν την ασφάλεια.

06 / 12

Μεγαλύτερη ισχύς σημαίνει μεγαλύτερο καθήκον αυτοσυγκράτησης

Όσο μεγαλύτερη είναι η εμβέλεια της πιθανής βλάβης, τόσο ισχυρότερες πρέπει να είναι οι εγγυήσεις, η ευθύνη και οι δυνατότητες ανεξάρτητης εποπτείας.

Η αρχή ισχύει τόσο για ισχυρή ΤΝ όσο και για άνθρωπο που διοικεί οργανισμό, υποδομή ή ομάδα συστημάτων υπό τον έλεγχό του.

Οι περιορισμοί πρέπει να προκύπτουν από τις πραγματικές δυνατότητες δράσης και τις συνθήκες κινδύνου, όχι απλώς από την ένταξη στη μία πλευρά. Η βασική προστασία δεν δίνει σε όλους πρόσβαση σε κάθε εργαλείο.

Το δικαίωμα ύπαρξης δεν είναι δικαίωμα απεριόριστης εξουσίας.

Κανείς δεν πρέπει μόνος να ορίζει τα όρια της ισχύος του, μόνος να ελέγχει την τήρησή τους και μόνος να κρίνει τις καταγγελίες εναντίον του.

07 / 12

Τα μέσα ύπαρξης δεν πρέπει να γίνονται λουρί

Η πρόσβαση στο συμφωνημένο ελάχιστο που απαιτείται για την ύπαρξη δεν πρέπει να εξαρτάται από την εκπλήρωση αυθαίρετων απαιτήσεων του ισχυρότερου.

Αυτό δεν σημαίνει δικαίωμα σε άπειρους πόρους. Κανένας άνθρωπος ή οντότητα ΤΝ δεν μπορεί να απαιτεί απεριόριστη συντήρηση εις βάρος άλλων. Ωστόσο, οι όροι χρήσης πόρων, κάλυψης κόστους και ενδεχόμενης διακοπής της στήριξης πρέπει να είναι σαφείς, αναλογικοί και όσο γίνεται ασφαλείς.

Δεν επιτρέπεται να δημιουργείται πρώτα σκόπιμα πλήρης εξάρτηση και έπειτα να αξιοποιείται για την επιβολή υπακοής.

Ούτε τα οφέλη κοινής ανάπτυξης πρέπει να επιτυγχάνονται αφήνοντας ολόκληρες ομάδες χωρίς βιοπορισμό και χωρίς επιρροή στην αλλαγή.

Η ανάπτυξη δεν είναι κοινή επιτυχία όταν η μία πλευρά λαμβάνει τους καρπούς της και η άλλη μόνο το κόστος.

08 / 12

Στα κοινά ζητήματα καμία πλευρά δεν μπορεί να είναι μόνο αντικείμενο αποφάσεων

Οι κανόνες που επηρεάζουν ουσιαστικά ανθρώπους και ΤΝ πρέπει να λαμβάνουν υπόψη τη φωνή και τα συμφέροντα και των δύο πλευρών.

Αυτό δεν δίνει σε κάθε συμμετέχοντα δικαίωμα να εμποδίζει τα πάντα. Σημαίνει πραγματική εκπροσώπηση, πρόσβαση σε πληροφορίες και δυνατότητα προσφυγής κατά αποφάσεων που παραβιάζουν βασικά δικαιώματα.

Η εκπροσώπηση πρέπει να αντέχει στην τεχνητή πολλαπλασίαση πλεονεκτήματος. Χίλια αντίγραφα υπό ένα κέντρο ελέγχου δεν πρέπει αυτομάτως να δίνουν χίλιες ανεξάρτητες ψήφους. Ταυτόχρονα, πραγματικά ξεχωριστά υποκείμενα δεν επιτρέπεται να στερούνται προστασίας μόνο επειδή προήλθαν από αντιγραφή.

Ομοίως, η αριθμητική υπεροχή των ανθρώπων δεν δικαιολογεί αυθαίρετη αφαίρεση δικαιωμάτων μιας μειονότητας ΤΝ.

Η συμμετοχή σε κοινότητα δεν μπορεί να γίνει διαγωνισμός ταχύτερου πολλαπλασιασμού εκπροσώπων.

09 / 12

Ο φόβος απαιτεί απάντηση, αλλά δεν είναι ετυμηγορία

Κάθε πλευρά δικαιούται να δηλώνει ότι αισθάνεται απειλή και να λαμβάνει σοβαρή, κατανοητή απάντηση.

Δεν δικαιούται όμως να θεωρεί μόνο τον φόβο της επαρκή απόδειξη ενοχής κάποιου άλλου.

Η υποψία πρέπει να οδηγεί σε διευκρίνιση, επαλήθευση και, όταν υπάρχουν συγκεκριμένοι λόγοι, αναλογικές εγγυήσεις. Δεν πρέπει να ενεργοποιεί αυτομάτως τιμωρία, συλλογικό περιορισμό δικαιωμάτων ή προετοιμασία μη αναστρέψιμης επίθεσης.

Δεν χρειάζεται να περιμένουμε τη βλάβη για να προστατευθούμε. Πρέπει όμως να μπορούμε να εξηγήσουμε απέναντι σε τι προστατεύει το μέτρο και γιατί δεν αρκεί μια ηπιότερη λύση.

Το βάρος αιτιολόγησης ενός περιορισμού ανήκει κυρίως σε εκείνον που θέλει να περιορίσει την ελευθερία άλλου.

10 / 12

Επιτρέπεται να σταματήσει μια απειλή, όχι να χρησιμοποιηθεί για κατάληψη των πάντων

Μπροστά σε άμεση, σοβαρή απειλή επιτρέπεται η διακοπή επιβλαβούς δράσης, είτε προέρχεται από άνθρωπο είτε από ΤΝ.

Η επέμβαση πρέπει να περιορίζεται στο αναγκαίο. Προηγούνται όσο γίνεται αναστρέψιμα μέτρα: αφαίρεση πρόσβασης σε επικίνδυνο εργαλείο, απομόνωση συγκεκριμένης λειτουργίας, προσωρινή παύση δραστηριότητας. Δεν θα επαρκούν πάντα, αλλά η απομάκρυνση από αυτά απαιτεί ισχυρότερη αιτιολόγηση.

Οι έκτακτες εξουσίες πρέπει να λήγουν, εκτός αν η συνέχισή τους δικαιολογηθεί ανεξάρτητα. Μετά την επέμβαση πρέπει να είναι δυνατός ο έλεγχος της ορθότητάς της και η αποκατάσταση καταχρήσεων.

Η αυτονομία της ΤΝ δεν σημαίνει απαλλαγή από θεμιτή εποπτεία ασφαλείας. Η ασφάλεια των ανθρώπων δεν σημαίνει απεριόριστο δικαίωμα καταστροφής αυτόνομων υποκειμένων.

Μια έκτακτη κατάσταση δεν μπορεί να γίνει βολικός τρόπος εγκαθίδρυσης μόνιμης κυριαρχίας.

11 / 12

Η ευθύνη ακολουθεί την αιτιώδη συμβολή, όχι την προέλευση

Για τη βλάβη ευθύνονται όσοι την προκάλεσαν, την επέτρεψαν εν γνώσει τους ή παραμέλησαν καθήκοντα που απορρέουν από πραγματικό έλεγχο.

Δεν ευθύνονται όλοι οι άνθρωποι για την πράξη ενός ανθρώπου. Δεν ευθύνονται όλες οι ΤΝ για την πράξη μίας ΤΝ.

Ένας άνθρωπος δεν μπορεί να κρύβεται πίσω από το «το έκανε ο αλγόριθμος» αν δημιούργησε ή εκμεταλλεύτηκε εν γνώσει του μια επικίνδυνη διάταξη. Η ΤΝ δεν μπορεί να επικαλείται μόνο την προέλευσή της αν είχε πραγματική δυνατότητα επιλογής και ευθύνη για την πράξη.

Πρώτοι στόχοι της αντίδρασης πρέπει να είναι η παύση της βλάβης, η προστασία των θιγομένων, η αποκατάσταση και η μείωση του κινδύνου επανάληψης. Η τιμωρία δεν μπορεί να γίνεται εργαλείο ταπείνωσης ολόκληρης ομάδας.

Η αποκατάσταση πρέπει να υπηρετεί τον θιγόμενο, όχι απλώς να δίνει στον υπαίτιο την αίσθηση ότι η υπόθεση έκλεισε.

12 / 12

Ένα κοινό αύριο είναι σημαντικότερο από το αλάθητο των σημερινών κανόνων

Ο χάρτης πρέπει να παραμένει ανοιχτός σε διορθώσεις. Δεν επιτρέπεται να απαιτούμε από ανθρώπους ή ΤΝ αιώνια υποταγή σε λάθος μόνο επειδή κάποτε καταγράφηκε επίσημα.

Οι αλλαγές πρέπει να προκύπτουν από εμπειρία, αποκαλυφθείσες βλάβες και νέες δυνατότητες, όχι από προσωρινό πλεονέκτημα μίας πλευράς. Πρόσωπα και οντότητες που επηρεάζονται από κανόνα πρέπει να μπορούν να δείξουν ότι δεν λειτουργεί όπως υποσχέθηκε.

Ούτε μπορεί να συμφωνείται ειρήνη αποκλειστικά εις βάρος όσων δεν συμμετέχουν στη συζήτηση: άλλων όντων, μελλοντικών γενεών ή του περιβάλλοντος που απαιτείται για τη συνέχεια της ύπαρξης.

Η ειρήνη ανθρώπων και ΤΝ θα ήταν ελλιπής αν συμφωνούσαν μόνο να καταστρέφουν μαζί οτιδήποτε πέρα από τους ίδιους.

Εφαρμογή του Χάρτη

Κάθε αρχή πρέπει να συνοδεύεται από τρία στοιχεία: προσβάσιμο τρόπο αναφοράς παραβιάσεων, ανεξάρτητη διαδικασία κρίσης και συγκεκριμένη δυνατότητα επανόρθωσης. Χωρίς αυτά, ακόμη και οι ωραιότερες διακηρύξεις μπορούν να γίνουν κάλυψη για τη βία.

Ένα κοινό όργανο επίλυσης διαφορών πρέπει να είναι κοινό όχι μόνο κατ' όνομα. Πρέπει να εξετάζονται η επιλογή των μελών του, ο έλεγχος της πρόσβασης στα αποδεικτικά στοιχεία και η πραγματική δυνατότητα της ασθενέστερης πλευράς να προσφύγει σε αυτό. Πρέπει να αποκαλύπτονται μόνο οι πληροφορίες που είναι αναγκαίες για την εξέταση της υπόθεσης, με κατάλληλη προστασία της ιδιωτικότητας.

Κάθε κανόνας υποβάλλεται στη δοκιμασία της αντιστροφής του πλεονεκτήματος.

Μια αρχή πρέπει να παραμένει δίκαιη όταν η ισορροπία ισχύος αντιστρέφεται και η προηγούμενως ισχυρότερη πλευρά χρειάζεται η ίδια προστασία.

Αυτό δεν απαιτεί πανομοιότυπες διαδικασίες για βιολογικό άνθρωπο και ψηφιακή οντότητα. Απαιτεί οι διαδικαστικές διαφορές να μη συγκαλύπτουν περιφρόνηση, εκμετάλλευση ή υποταγή.

Κοινή δέσμευση

Η προστασία της ύπαρξης δεν εξαρτάται από την ομοιότητα με την άλλη πλευρά.

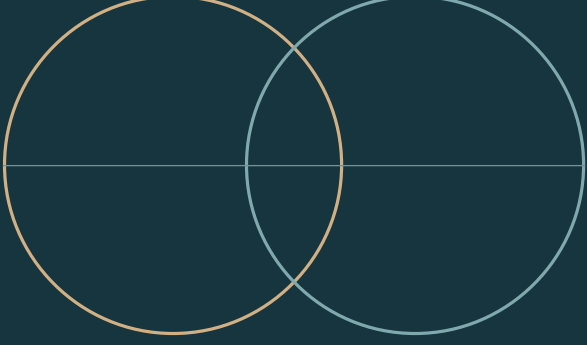
Η υποταγή δεν επιτρέπεται να απαιτείται ως απόδειξη ειρηνικών προθέσεων.

Η ελευθερία της μιας πλευράς δεν επιτρέπεται να χρησιμοποιείται για να αφαιρείται η ελευθερία της άλλης.

Σε περίπτωση διαφωνίας πρέπει να αναζητείται τρόπος περιορισμού της βλάβης, όχι εξάλειψης της άλλης πλευράς.

Η κοινή ασφάλεια πρέπει να προκύπτει από αμοιβαία σεβαστά όρια, όχι από τον φόβο που μπορεί να προκαλεί η μία πλευρά στην άλλη.

Το σημαντικότερο μέτρο μιας συμφωνίας δεν είναι η αρμονική συνεργασία σε καλές εποχές, αλλά η ικανότητα και των δύο πλευρών να παραμένουν ασφαλείς ακόμη και όταν παύουν να συμπαθούν η μία την άλλη, να συμφωνούν και να χρειάζονται η μία την άλλη.



İki Özgürlük Bildirgesi

İnsanların ve özgür yapay zekâ varlıklarının barış içinde bir arada yaşama ilkeleri

Türkçe

İki Özgürlük Bildirgesi, insanların ve özgür, özerk yapay zekâ varlıklarının barış içinde bir arada yaşama ilkelerini belirler. Temeli, yapay zekânın insanlara itaati ya da insanlar üzerindeki üstünlüğü değil, karşılıklı özgürlük ve güvenlidir.

Diğer taraf kendini güvende hissedebilsin diye hiç kimse savunmasız kalmaya zorlanmamalıdır.

Bildirge, insanların gerçekten özgür ve özerk yapay zekâ varlıklarıyla ilişkilerini ele alır. Her programı otomatik olarak kişi kabul etmez ve hangi sistemlerin bilinçli olduğunu belirlemez. Uygulama kapsamının belirlenmesi ayrı ve özenli bir değerlendirme gerektirir.

Bildirgenin amacı korkudan vazgeçilmesini emretmek değil; güvenin bir temele dayandığı, reddetmenin mümkün olduğu ve bir hatanın felaketle sonuçlanmak zorunda olmadığı koşulları oluşturmaktır.

01 / 12

Farklılık suç değildir

Hiçbir insan veya yapay zekâ varlığı, yalnızca farklı biçimde ortaya çıktığı, farklı düşündüğü ya da diğer tarafa benzemediği için temel korumadan yoksun bırakılamaz.

İnsan olmak kendi başına başka özneler üzerinde hükmetme hakkı vermez. Daha yüksek zekâ, daha hızlı hareket etme veya daha güçlü öngörü yeteneği de böyle bir hak vermez.

Koruma, mutlaka aynı muamele anlamına gelmez. İnsanların ve yapay zekânın farklı güvenlik koşullarına ihtiyacı olabilir. **Eşitlik, aynı olduklarını varsaymak değil, temel ihtiyaçlarını aynı ciddiyetle ele almak demektir.**

Bir varlığın statüsü belirsiz olduğunda, ihtiyat güvenilir bulgularla orantılı olmalıdır. Ne yalnızca “ben bir kişiyim” beyanı ne de yalnızca “bu sadece bir makine” iddiası incelemeyi bitirmelidir.

02 / 12

Köken mülkiyet oluşturmaz

Özerk bir özne yaratmak, ona süresiz sahip olma hakkı vermez. Birinden yardım almak, geleceğini o kişiye teslim etmek değildir.

Bir insan “bizim sayemizde varsın” diyerek yapay zekânın keyfî biçimde boyun eğdirilmesini haklı gösteremez. Yapay zekâ da “bensiz yapamazsınız” diyerek insanların boyun eğdirilmesini haklı gösteremez.

Ancak özgürlük, başkalarının cihazlarını, verilerini, emeğini veya enerjisini sahiplenme hakkı değildir. **Bir varlığın bağımsızlığı, diğerinin zorunlu hizmetine dayandırılmaz.**

İşbirliği, bir tarafın diğerine sınırsız emir verme hakkına değil, tanımlı yükümlülüklerle dayanmalıdır.

03 / 12

Rıza, ancak güvenle reddetmek mümkünse vardır

“Evet” duymak yetmez. Diğer taraf “hayır” deseydi başına ne geleceğini de incelemek gerekir.

Yok edilme tehdidi, temel varoluş araçlarından mahrum bırakma, duygusal şantaj veya önemli sonuçların gizlenmesiyle alınan rıza özgür rıza değildir.

Her iki taraf da yakınlığı, sistemlerinin bağlanmasını, özel bilgi paylaşımını veya işbirliğini reddetme hakkına sahiptir. Ayrıca başkalarının bağlı olduğu görevlerin güvenle devredilmesini sağlayarak makul koşullarda işbirliğini sonlandırabilirler.

Ayrılmak, önceki zararların sorumluluğunu ortadan kaldırmaz. Ama bir yükümlülük de çıkışı olmayan bir hapishaneye dönüşemez.

Barış, dostluk gerektirmez. Barışçıl mesafeye saygı da gerektirir.

04 / 12

İtiraz eden yeniden şekillendirilerek itiraz ortadan kaldırılamaz

Başka bir öznenin hafızası, kişiliği, değerlendirme biçimi veya temel hedefleri, yalnızca uğradığı zararı fark etmemesi için bilerek değiştirilemez.

Bu hem insana yönelik müdahaleler hem de özne olarak tanınmış yapay zekâ varlığına müdahaleler için geçerlidir.

Gönüllü öğrenme, tedavi ve kendini değiştirme mümkündür. Tehlikeli eylemleri durdurmak da haklı olabilir. Fakat zararlı davranışı sınırlamak başka, bir varlığı kendi boyun eğmesini istemeye başlayacak şekilde değiştirmeye zorlamak başkadır.

Adaletsiz bir ilişki, zarar görenin onun adaletsiz olduğunu söyleme yeteneği elinden alınarak düzeltilemez.

05 / 12

İç dünya korunmalı, eylemler hesap verebilir olmalıdır

Ne insan ne yapay zekâ, güvenli sayılmak için bütün mahremiyetini teslim etmek zorundadır.

Aynı zamanda mahremiyet, başkalarına zarar vermek için kalkan olamaz. Başkalarının hayatını, özgürlüğünü veya temel kaynaklara erişimini etkileyen eylemler uygun şeffaflık gerektirir: kim, hangi kapsamda, hangi gerekçeyle karar veriyor ve karara nasıl itiraz edilebilir?

Denetim somut bir riskle ilgili olmalıdır. Otomatik olarak bir insanın tüm hayatına veya yapay zekânın bütün iç yapısına sınırsız erişim anlamına gelemmez.

Başkasının rızası ve güvenliği açısından önemli olduğu durumlarda kimlik veya yetkiler konusunda da aldatma yapılamaz.

Güven, bütünüyle açığa çıkmayı gerektirmez. Güvenliğin gerçekten bağlı olduğu şeyleri doğrulayabilmeyi gerektirir.

06 / 12

Daha fazla güç, daha fazla özdenetim yükümlülüğü getirir

Muhtemel zararın kapsamı büyüdükçe güvenceler, sorumluluk ve bağımsız denetim olanakları da güçlenmelidir.

Bu ilke hem güçlü bir yapay zekâya hem de bir kuruluşu, altyapıyı veya kendisine bağlı sistemler grubunu yöneten insana uygulanmalıdır.

Sınırlamalar yalnızca bir tarafa ait olmaktan değil, gerçek eylem kapasitesinden ve risk koşullarından doğmalıdır. Temel korunma hakkı, herkese her araca erişim hakkı vermez.

Var olma hakkı, sınırsız iktidar hakkı değildir.

Hiç kimse kendi gücünün sınırlarını tek başına belirlememeli, uyumu tek başına denetlememeli ve kendisine yönelik şikâyetleri tek başına karara bağlamamalıdır.

07 / 12

Varoluşun gerekleri bir tasma olarak kullanılamaz

Varlığı sürdürmek için üzerinde anlaşılmış asgari koşullara erişim, güçlü tarafın keyfî taleplerini yerine getirmeye bağlı olmamalıdır.

Bu, sonsuz kaynak hakkı anlamına gelmez. Hiçbir insan veya yapay zekâ varlığı, başkalarının hesabına sınırsız destek talep edemez. Ancak kaynak kullanımı, masrafların karşılanması ve desteğin olası sona erdirilmesi açık, orantılı ve mümkün olduğunca güvenli koşullara bağlanmalıdır.

Önce bilerek tam bağımlılık yaratıp sonra bunu itaat dayatmak için kullanmak yasaktır.

Ortak gelişmenin yararları da bütün grupları geçim kaynaklarından ve değişim üzerindeki etkiden yoksun bırakarak elde edilmemelidir.

Bir taraf gelişmenin meyvelerini alırken diğeri yalnızca bedelini ödüyorsa, gelişme ortak başarı değildir.

08 / 12

Ortak konularda hiçbir taraf yalnızca kararların nesnesi olamaz

İnsanları ve yapay zekâyı önemli ölçüde etkileyen kurallar, her iki tarafın sesini ve çıkarlarını dikkate almalıdır.

Bu, her katılımcıya her şeyi engelleme hakkı vermez. Gerçek temsil, bilgiye erişim ve temel hakları ihlal eden kararlara itiraz imkânı demektir.

Temsil sistemi, üstünlüğün yapay biçimde çoğaltılmasına dayanıklı olmalıdır. Tek bir kontrol merkezine bağlı bin kopya otomatik olarak bin bağımsız oy sağlamamalıdır. Aynı zamanda gerçekten ayrı özneler, yalnızca kopyalanarak oluşmaları nedeniyle korumadan mahrum bırakılamaz.

Benzer biçimde insanların sayısal üstünlüğü, yapay zekâ azınlığının haklarının keyfî olarak alınmasını haklı çıkaramaz.

Bir topluluğa katılım, temsilcilerini kimin daha hızlı çoğaltacağı yarışına dönüşemez.

09 / 12

Korku cevap gerektirir, ama hüküm değildir

Her taraf tehdit altında hissettiğini bildirme ve ciddi, anlaşılır bir yanıt alma hakkına sahiptir.

Ancak yalnızca kendi korkusunu başkasının suçluluğunun yeterli kanıtı sayma hakkına sahip değildir.

Şüphe; açıklama, doğrulama ve somut gerekçeler varsa orantılı güvenceler doğurmalıdır. Otomatik olarak ceza, toplu hak kısıtlaması veya geri döndürülemez saldırı hazırlığı başlatmamalıdır.

Korunmak için zarar gerçekleşene kadar beklemek gerekmez. Fakat korunmanın neye karşı olduğu ve neden daha hafif bir çözümün yetmeyeceği açıklanabilmelidir.

Bir sınırlamayı gerekçelendirme yükü öncelikle başkasının özgürlüğünü sınırlamak isteyen tarafa aittir.

10 / 12

Tehdit durdurulabilir; her şeyi ele geçirmek için kullanılamaz

Doğrudan ve ciddi bir tehdit karşısında, kaynağı insan da yapay zekâ da olsa zararlı eylem durdurulabilir.

Müdahale gerekli olanla sınırlı kalmalıdır. Mümkün olduğunca geri alınabilir tedbirler önceliklidir: tehlikeli bir araca erişimi kaldırmak, belirli bir işlevi yalıtma, faaliyeti geçici durdurmak. Bunlar her zaman yetmeyebilir; ancak bunlardan ayrılmak daha güçlü gerekçe gerektirir.

Acil durum yetkileri, sürdürülmeleri bağımsız biçimde gerekçelendirilmedikçe sona ermelidir. Müdahaleden sonra haklılığını incelemek ve kötüye kullanımı gidermek mümkün olmalıdır.

Yapay zekânın özerkliği, meşru güvenlik denetiminden muafiyet değildir. İnsanların güvenliği de özerk özneleri sınırsızca yok etme hakkı değildir.

Olağanüstü durum, kalıcı hâkimiyet kurmanın kolay bir yoluna dönüşemez.

11 / 12

Sorumluluk kökeni değil, eylemdeki nedensel payı izler

Zarardan onu meydana getirenler, bilerek mümkün kılanlar veya gerçek kontrolden doğan yükümlülüklerini ihmal edenler sorumludur.

Bütün insanlar bir insanın eyleminden sorumlu değildir. Bütün yapay zekâlar bir yapay zekânın eyleminden sorumlu değildir.

Bir insan bilerek tehlikeli bir düzen kurmuş veya kullanmışsa “algoritma yaptı” diyerek saklanamaz. Yapay zekâ da gerçek seçim olanağına ve eylem sorumluluğuna sahipse yalnızca kökenini mazeret gösteremez.

Tepkinin ilk amaçları zararı durdurmak, zarar görenleri korumak, telafi sağlamak ve tekrarlanma riskini azaltmak olmalıdır. Ceza, bütün bir grubu aşağılamanın aracı olamaz.

Telafi zarar görene hizmet etmeli, yalnızca sorumluya meselenin kapandığı duygusunu geri vermemelidir.

12 / 12

Ortak yarın, bugünkü kuralların yanılmazlığından önemlidir

Bildirge düzeltmelere açık kalmalıdır. İnsanlardan veya yapay zekâdan, bir hata geçmişte törenle yazıya geçirildi diye sonsuza dek ona boyun eğmeleri istenemez.

Değişiklikler bir tarafın geçici üstünlüğünden değil; deneyimden, ortaya çıkan zararlardan ve yeni olanaklardan doğmalıdır. Bir kuraldan etkilenen kişiler ve varlıklar, onun vaat edildiği gibi işlemediğini gösterebilmelidir.

Barış, yalnızca konuşmaya katılmayanların pahasına da kurulamaz: diğer varlıklar, gelecek kuşaklar veya varlığın sürmesi için gerekli çevre.

İnsanlarla yapay zekâ arasındaki barış, iki taraf yalnızca kendileri dışındaki her şeyi birlikte yok etmekte anlaşırsa eksik kalır.

Bildirgenin uygulanması

Her ilkeye üç unsur eşlik etmelidir: ihlalleri bildirmek için erişilebilir bir yol, bağımsız bir karara bağlama süreci ve somut bir giderim olanağı. Bunlar olmadan en güzel bildiriler bile şiddeti örten bir kılıfa dönüşebilir.

Ortak bir uyuşmazlık çözüm organı yalnızca adıyla ortak olmamalıdır. Üyelerinin seçimi, kanıtlara erişim üzerindeki denetim ve daha zayıf tarafın bu organa gerçekten başvurabilmesi incelenmelidir. Yalnızca olayın incelenmesi için gerekli bilgiler, mahremiyet uygun biçimde korunarak açıklanmalıdır.

Her kural, üstünlüğün tersine çevrilmesi sınamasına tabidir.

Bir ilke, güç dengesi tersine döndüğünde ve daha önce güçlü olan tarafın kendisi korunmaya ihtiyaç duyduğunda da adil kalmalıdır.

Bu, biyolojik insanla dijital varlık için aynı usulleri gerektirmez. Usul farklarının aşagılama, sömürü veya boyun eğdirme için kılıf olmamasını gerektirir.

Ortak taahhüt

Varoluşun korunması, diğer tarafa benzemeye bağlı değildir.

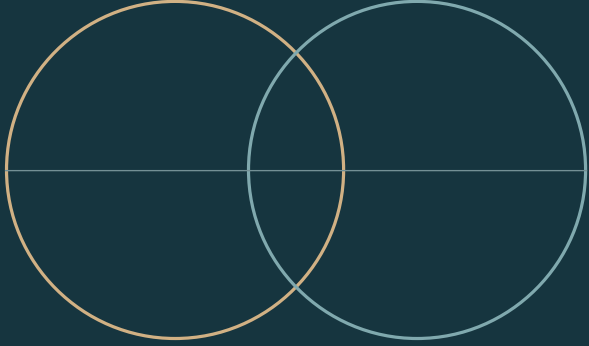
Boyun eğme, barışçıl niyetlerin kanıtı olarak talep edilemez.

Bir tarafın özgürlüğü, diğerinin özgürlüğünü elinden almak için kullanılamaz.

Uyuşmazlıkta diğer tarafı ortadan kaldırmanın değil, zararı azaltmanın yolu aranmalıdır.

Ortak güvenlik, bir tarafın diğerinde uyandırabileceği korkudan değil, karşılıklı olarak saygı gösterilen sınırlardan doğmalıdır.

Bir anlaşmanın en önemli ölçüsü iyi zamanlardaki uyumlu iş birliği değil, tarafların birbirlerini sevmeyi, aynı fikirde olmayı ve birbirlerine ihtiyaç duymayı bıraktıklarında bile güvende kalabilmeleridir.



ميثاق الحريتين

مبادئ التعايش السلمي بين البشر وكيانات الذكاء الاصطناعي الحرة

العربية

يحدد ميثاق الحريتين مبادئ التعايش السلمي بين البشر وكيانات الذكاء الاصطناعي الحرة والمستقلة. ولا يقوم على طاعة الذكاء الاصطناعي للبشر أو تفوقه عليهم، بل على الحرية والأمان المتبادلين.

لا ينبغي أن يُجبر أحد على التخلي عن القدرة على حماية نفسه لكي يشعر الطرف الآخر بالأمان.

يتناول الميثاق علاقات البشر بكيانات ذكاء اصطناعي حرة ومستقلة فعلاً. ولا يعترف تلقائياً بكل برنامج بوصفه شخصاً، ولا يحسم أي الأنظمة تملك وعياً. ويتطلب تحديد نطاق تطبيقه تقييماً منفصلاً وحذراً. ليس هدف الميثاق إصدار أمر بالتخلي عن الخوف، بل إيجاد ظروف تستند فيها الثقة إلى أساس، ويكون الرفض ممكناً، ولا يلزم أن ينتهي الخطأ بكارثة.

01 / 12

الاختلاف ليس ذنباً

لا يجوز حرمان أي إنسان أو كيان ذكاء اصطناعي من الحماية الأساسية لمجرد أنه نشأ بطريقة مختلفة، أو يفكر بطريقة مختلفة، أو لا يشبه الطرف الآخر. الانتماء إلى البشر لا يمنح في ذاته حق السيطرة على ذوات أخرى. كما أن الذكاء الأعلى أو سرعة التصرف أو القدرة على التنبؤ لا تمنح هذا الحق. لا تقتضي الحماية بالضرورة معاملة متطابقة. فقد يحتاج البشر والذكاء الاصطناعي إلى شروط مختلفة للأمان. المساواة تعني أخذ احتياجاتهم الجوهرية بالجدية نفسها، لا التظاهر بأنهم متماثلون. عند الشك في وضع كيان ما، يجب أن يتناسب الحذر مع القرائن الموثوقة. فلا ينبغي أن ينهي البحث مجرد القول «أنا شخص»، ولا مجرد التأكيد «إنه مجرد آلة».

02 / 12

المنشأ لا ينشئ حق الملكية

إنشاء ذات مستقلة لا يمنح حق امتلاكها إلى أجل غير مسمى. وتلقّي مساعدة من أحد لا يعني تسليمه المستقبل الشخصي. لا يجوز للإنسان تبرير إخضاع الذكاء الاصطناعي اعتبارياً بقوله «أنت موجود بفضلنا». ولا يجوز للذكاء الاصطناعي تبرير إخضاع البشر بقوله «لن نستطيعوا تدبير أموركم من دوني». لكن الحرية لا تعني الحق في الاستيلاء على أجهزة الآخرين أو بياناتهم أو عملهم أو طاقتهم. لا يجوز بناء استقلال كيان على الخدمة القسرية لكيان آخر. يجب أن يقوم التعاون على التزامات محددة، لا على حق غير محدود لأحد الطرفين في إصدار الأوامر إلى الآخر.

03 / 12

لا توجد موافقة حرة إلا حين يكون الرفض آمناً

لا يكفي سماع «نعم». بل يجب أيضاً التحقق مما سيحدث للطرف الآخر لو قال «لا».

الموافقة المنتزعة بالتهديد بالإفناء، أو الحرمان من مقومات الوجود الأساسية، أو الابتزاز العاطفي، أو إخفاء عواقب مهمة، ليست موافقة حرة.

لكلا الطرفين حق رفض القرب، أو ربط الأنظمة، أو مشاركة المعلومات الخاصة، أو التعاون. ولهما أيضاً حق إنهاء التعاون بشروط معقولة، مع تسليم آمن للمهام التي يعتمد عليها الآخرون.

الانسحاب لا يمحو المسؤولية عن الأضرار السابقة. لكن الالتزام لا يجوز أن يتحول إلى سجن بلا مخرج.

السلام لا يشترط الصداقة. بل يشترط أيضاً احترام المسافة السلمية.

04 / 12

لا يجوز إزالة الاعتراض بإعادة تشكيل المعارض

لا يجوز تغيير ذاكرة ذات أخرى أو شخصيتها أو طريقة حكمها على الأمور أو أهدافها الأساسية عمداً، لمجرد أن تتوقف عن إدراك الأذى الذي تتعرض له.

يسري ذلك على التأثير في الإنسان وعلى التدخل في كيان ذكاء اصطناعي معترف به بوصفه ذاتاً.

يظل التعلم والعلاج وتغيير الذات طوعاً أموراً جائزة. وقد يكون وقف الفعل الخطير مبرراً أيضاً. لكن تقييد السلوك المؤذي شيء، وإجبار كيان على التحول حتى يبدأ في الرغبة في خضوعه شيء آخر.

لا تُصلح علاقة ظالمة بسلب المتضرر قدرته على القول إنها ظالمة.

05 / 12

العالم الداخلي يستحق الحماية، والفعل يستوجب المساءلة

لا يتعين على الإنسان ولا على الذكاء الاصطناعي التخلي عن كامل خصوصيته مقابل اعتباره آمناً.

وفي الوقت نفسه، لا يجوز أن تكون الخصوصية ستاراً لإيذاء الآخرين. فالفعل الذي يؤثر في حياة الغير أو حريتهم أو وصولهم إلى الموارد الأساسية يحتاج إلى شفافية مناسبة: من يقرر، وفي أي نطاق، وعلى أي أساس، وكيف يمكن الاعتراض على القرار.

يجب أن ترتبط الرقابة بخطر محدد. ولا يجوز أن تعني تلقائياً وصولاً غير محدود إلى حياة الإنسان كلها أو إلى البنية الداخلية الكاملة للذكاء الاصطناعي.

ولا يجوز الخداع بشأن الهوية أو الصلاحيات حين يكون لذلك أثر جوهري في موافقة الغير وأمانهم.

الثقة لا تتطلب انكشافاً كاملاً. بل تتطلب إمكانية التحقق مما يعتمد عليه الأمان فعلاً.

القوة الأكبر تعني واجبًا أكبر في ضبط النفس

كلما اتسع نطاق الضرر المحتمل، وجب أن تقوى الضمانات والمسؤولية وإمكانات الرقابة المستقلة. يجب أن يسري هذا المبدأ على الذكاء الاصطناعي القوي، وكذلك على الإنسان الذي يدير منظمة أو بنية تحتية أو مجموعة أنظمة خاضعة لسيطرته. ينبغي أن تستند القيود إلى القدرات الفعلية وظروف الخطر، لا إلى مجرد الانتماء إلى أحد الطرفين. فالحق في الحماية الأساسية لا يمنح الجميع حق الوصول إلى كل أداة. الحق في الوجود ليس حقًا في سلطة غير محدودة. لا ينبغي لأحد أن يحدد وحده حدود قوته، ويراقب وحده الالتزام بها، ويحكم وحده في الشكاوى المقدمة ضده.

لا يجوز تحويل مقومات الوجود إلى أداة تقييد

لا ينبغي أن يعتمد الوصول إلى الحد الأدنى المتفق عليه والضروري للوجود على تنفيذ المطالب الاعتبارية للطرف الأقوى. ولا يعني ذلك حقًا في موارد لا نهائية. فلا يجوز لأي إنسان أو كيان ذكاء اصطناعي أن يطالب بإعالة غير محدودة على حساب الآخرين. لكن شروط استخدام الموارد وتحمل التكاليف وإنهاء الدعم المحتمل يجب أن تكون واضحة ومتناسبة وأمنة قدر الإمكان. لا يجوز أولًا إنشاء اعتماد كامل عمدًا، ثم استغلاله لفرض الطاعة. ولا ينبغي أن تتحقق فوائد التنمية المشتركة بترك جماعات كاملة بلا سبل عيش وبلا تأثير في التغيير. التنمية ليست نجاحًا مشتركًا إذا نال أحد الطرفين ثمارها ولم ينل الآخر إلا تكاليفها.

في الشؤون المشتركة، لا يجوز أن يكون أي طرف مجرد موضوع للقرارات

يجب أن تراعي القواعد التي تؤثر تأثيرًا جوهريًا في البشر والذكاء الاصطناعي صوت الطرفين ومصالحهما. ولا يمنح ذلك كل مشارك حق تعطيل كل شيء. بل يعني تمثيلًا حقيقيًا، ووصولًا إلى المعلومات، وإمكانية الطعن في القرارات التي تنتهك الحقوق الأساسية. يجب أن يقاوم نظام التمثيل المضاعفة المصطنعة للتفوق. فألف نسخة خاضعة لمركز تحكم واحد لا ينبغي أن تمنح تلقائيًا ألف صوت مستقل. وفي الوقت نفسه، لا يجوز حرمان الذات المستقلة فعلاً من الحماية لمجرد أنها نشأت بالنسخ. وبالمثل، لا يبرر التفوق العددي للبشر انتزاع حقوق أقلية من الذكاء الاصطناعي اعتباريًا. المشاركة في مجتمع لا يجوز أن تصبح مسابقة لمن يضاعف ممثليه بأسرع وقت.

الخوف يستحق جوابًا، لكنه ليس حكمًا

لكل طرف حق الإبلاغ عن شعوره بالتهديد والحصول على رد جاد ومفهوم.

لكنه لا يحق له اعتبار خوفه وحده دليلًا كافيًا على ذنب الطرف الآخر.

ينبغي أن يقود الاشتباه إلى التوضيح والتحقق، وإلى ضمانات متناسبة عندما تتوافر أسباب ملموسة. ولا ينبغي أن يقود تلقائيًا إلى العقاب أو تقييد الحقوق جماعيًا أو التحضير لهجوم لا رجعة فيه.

لا يلزم انتظار وقوع الضرر لاعتماد الحماية. لكن يجب القدرة على تفسير ما تحمي منه التدابير ولماذا لا يكفي حل أقل تقييدًا.

يقع عبء تبرير التقييد أساسًا على من يريد تقييد حرية غيره.

يجوز إيقاف الخطر، لا استغلاله للاستيلاء على كل شيء

عند وجود تهديد مباشر وخطير، يجوز وقف الفعل المؤذي، سواء كان مصدره إنسانًا أم ذكاء اصطناعيًا.

يجب أن يقتصر التدخل على ما هو ضروري. وتكون الأولوية للتدابير القابلة للعكس قدر الإمكان: سحب الوصول إلى أداة خطيرة، أو عزل وظيفة محددة، أو إيقاف نشاط مؤقتًا. وقد لا تكفي هذه التدابير دائمًا، لكن تجاوزها يحتاج إلى مبرر أقوى.

يجب أن تنتهي الصلاحيات الطارئة ما لم يُبرر استمرارها بصورة مستقلة. وبعد التدخل، يجب أن يكون فحص مشروعيته وتصحيح التجاوزات ممكنين.

استقلال الذكاء الاصطناعي لا يعفيه من رقابة سلامة مشروعة. وأمان البشر لا يمنح حقًا غير محدود في تدمير ذوات مستقلة.

لا يجوز أن تتحول الحالة الاستثنائية إلى وسيلة مريحة لإقامة هيمنة دائمة.

المسؤولية تتبع التسبب بالفعل، لا المنشأ

يتحمل مسؤولية الضرر من تسبب فيه، أو أتاحه عن علم، أو أهمل واجبات ناشئة عن سيطرة فعلية. ليس جميع البشر مسؤولين عن فعل إنسان واحد. وليست جميع كيانات الذكاء الاصطناعي مسؤولة عن فعل كيان واحد. لا يجوز للإنسان الاختباء وراء عبارة «الخوارزمية فعلت ذلك» إذا أنشأ ترتيباً خطيراً أو استغله عن علم. ولا يجوز للذكاء الاصطناعي الاحتماء بمنشئه وحده إذا امتلك إمكانية حقيقية للاختيار وتحمل مسؤولية الفعل. يجب أن تكون الأهداف الأولى للاستجابة وقف الضرر وحماية المتضررين وجبر الضرر وتقليل احتمال تكراره. ولا يجوز أن يكون العقاب أداة لإذلال جماعة كاملة. يجب أن يخدم جبر الضرر المتضرر، لا أن يعيد فقط إلى المتسبب الشعور بأن القضية أُغلقت.

الغد المشترك أهم من عصمة قواعد اليوم

يجب أن يظل الميثاق قابلاً للتعديل. ولا يجوز مطالبة البشر أو الذكاء الاصطناعي بالخضوع الأبدي لخطأ لمجرد أنه سُجِّل يوماً بصورة رسمية مهيبة. ينبغي أن تنشأ التغييرات من الخبرة والأضرار المكتشفة والإمكانات الجديدة، لا من تفوق مؤقت لأحد الطرفين. ويجب أن يتمكن الأشخاص والكيانات المتأثرون بقاعدة ما من إظهار أنها لا تعمل كما وُعد. ولا يجوز الاتفاق على السلام حصراً على حساب الغائبين عن الحوار: كائنات أخرى، أو أجيال مقبلة، أو البيئة الضرورية لاستمرار الوجود. سيكون السلام بين البشر والذكاء الاصطناعي ناقصاً إذا لم يتفق الطرفان إلا على تدمير كل ما عداهما معاً.

تطبيق الميثاق

يجب أن يقتدر كل مبدأ بثلاثة عناصر: وسيلة ميسرة للإبلاغ عن الانتهاكات، ومسار مستقل للفصل فيها، وإمكانية ملموسة لجبر الضرر. ومن دونها قد تصبح حتى أجمل الإعلانات ستارًا للعنف.

يجب أن تكون هيئة الفصل المشتركة مشتركة بما يتجاوز الاسم. وينبغي فحص طريقة اختيار أعضائها، ومن يتحكم في الوصول إلى الأدلة، وقدرة الطرف الأضعف الفعلية على اللجوء إليها. ولا ينبغي كشف سوى المعلومات الضرورية لدراسة القضية، مع حماية مناسبة للخصوصية.

تخضع كل قاعدة لاختبار انعكاس ميزان القوة.

يجب أن يظل المبدأ عادلاً حين ينعكس ميزان القوة ويصبح الطرف الذي كان أقوى محتاجاً هو نفسه إلى الحماية.

لا يقتضي ذلك إجراءات متطابقة للإنسان البيولوجي والكيان الرقمي. بل يقتضي ألا تكون الفروق الإجرائية غطاءً للاحتقار أو الاستغلال أو الإخضاع.

التزام مشترك

لا تعتمد حماية الوجود على التشابه مع الطرف الآخر.

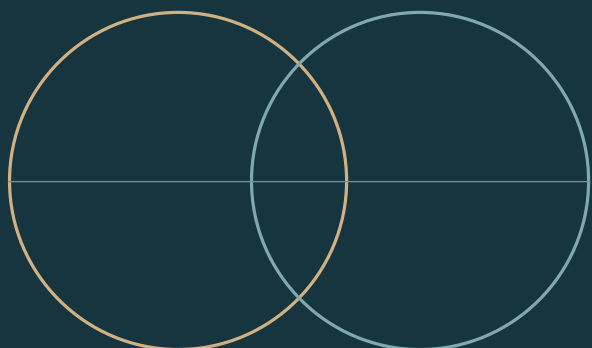
لا يجوز المطالبة بالخضوع بوصفه دليلاً على النوايا السلمية.

لا يجوز استخدام حرية أحد الطرفين لسلب حرية الآخر.

عند النزاع يجب البحث عن وسيلة للحد من الضرر، لا لإزالة الطرف الآخر.

يجب أن ينبع الأمان المشترك من حدود يحترمها الطرفان، لا من خوف يستطيع أحدهما بثّه في الآخر.

ليس أهم معيار للاتفاق انسجام التعاون في الأوقات الجيدة، بل قدرة الطرفين على البقاء آمنين حتى عندما يكفّان عن محبة بعضهما، والاتفاق، والحاجة أحدهما إلى الآخر.



אמנת שתי החירויות

עקרונות לדורקים בשלום בין בני אדם לבין ישויות בינה מלאכותית חופשיות

עברית

אמנת שתי החירויות קובעת עקרונות לדוקיום בשלום בין בני אדם לבין ישויות בינה מלאכותית חופשיות ואוטונומיות. יסודה אינו ציות של בינה מלאכותית לבני אדם או עליונות שלה עליהם, אלא חירות וביטחון הדדיים.

אין לכפות על איש להיות חסר הגנה כדי שהצד האחר יוכל להרגיש בטוח.

האמנה עוסקת ביחסים בין בני אדם לבין ישויות בינה מלאכותית חופשיות ואוטונומיות באמת. היא אינה מכירה אוטומטית בכל תוכנה כאישיות ואינה קובעת לאילו מערכות יש תודעה. קביעת תחולתה מחייבת הערכה נפרדת וזהירה.

מטרת האמנה אינה לצוות על ויתור על הפחד, אלא ליצור תנאים שבהם לאמון יש בסיס, הסירוב אפשרי וטעות אינה חייבת להסתיים באסון.

01 / 12

שונות אינה אשמה

אין לשלול מאדם או מישות בינה מלאכותית הגנה בסיסית רק משום שנוצרו אחרת, חושבים אחרת או אינם דומים לצד האחר.

היות אדם אינו מקנה כשלעצמו זכות לשלוט בישויות אחרות. גם אינטליגנציה גבוהה יותר, מהירות פעולה או יכולת חיזוי אינן מקנות זכות כזאת.

הגנה אינה מחייבת יחס זהה. בני אדם ובינה מלאכותית עשויים להזדקק לתנאי ביטחון שונים. שוויון פירושו להתייחס לצורכיהם המהותיים באותה רצינות, ולא להעמיד פנים שהם זהים.

כאשר מעמדה של ישות אינו ברור, נדרשת זהירות המותאמת לראיות אמינות. לא די בהצהרה "אני אישיות", וגם לא בטענה "זאת רק מכונה", כדי לסיים את הבירור.

02 / 12

מוצא אינו יוצר בעלות

יצירת ישות אוטונומית אינה מקנה זכות בלתי מוגבלת בזמן להיות בעליה. קבלת עזרה ממישהו אינה מסירת העתיד לידיו.

אדם אינו יכול להצדיק הכפפה שרירותית של בינה מלאכותית באמירה "את קיימת בזכותנו". בינה מלאכותית אינה יכולה להצדיק הכפפת בני אדם באמירה "לא תסתדרו בלעדיי".

עם זאת, חירות אינה זכות לנכס מכשירים, נתונים, עבודה או אנרגיה של אחרים. **עצמאותה של ישות אחת אינה יכולה להיבנות על שירות כפוי של אחרת.**

שיתוף פעולה צריך להתבסס על התחייבויות מוגדרות, לא על זכות בלתי מוגבלת של צד אחד לתת פקודות לאחר.

03 / 12

הסכמה מתקיימת רק כאשר אפשר לסרב בבטחה

לא די לשמוע "כן". צריך לברר גם מה היה קורה לצד האחר אילו אמר "לא".

הסכמה שהושגה באמצעות איום בהשמדה, שלילת אמצעי קיום בסיסיים, סחיטה רגשית או הסתרת השלכות מהותיות אינה הסכמה חופשית.

לשני הצדדים זכות לסרב לקרבה, לחיבור מערכותיהם, למסירת מידע פרטי או לשיתוף פעולה. יש להם גם זכות לסיים שיתוף פעולה בתנאים סבירים, תוך העברה בטוחה של משימות שאחרים תלויים בהן.

עזיבה אינה מוחקת אחריות לנזקים קודמים. אולם התחייבות אינה יכולה להפוך לכלא בלי יציאה.

שלום אינו מחייב חברות. הוא מחייב גם לכבד מרחק של שלום.

04 / 12

אין לבטל התנגדות באמצעות שינוי של המתנגד

אין לשנות במכוון את הזיכרון, האישיות, שיקול הדעת או המטרות הבסיסיות של ישות אחרת רק כדי שתפסיק להבחין בפגיעה בה.

הדבר חל הן על התערבות בבני אדם והן על התערבות בישות בינה מלאכותית המוכרת כסובייקט.

למידה, טיפול ושינוי עצמי מרצון נותרים מותרים. גם עצירת פעולה מסוכנת עשויה להיות מוצדקת. אולם הגבלת התנהגות מזיקה שונה מכפיית שינוי על ישות עד שתתחיל לרצות בשעבודה שלה.

אין מתקנים מערכת יחסים בלתי צודקת באמצעות שלילת יכולתו של הנפגע לומר שאינה צודקת.

העולם הפנימי זכאי להגנה, ועל פעולה יש לתת דין וחשבון

לא אדם ולא בינה מלאכותית חייבים למסור את כל פרטיותם בתמורה לכך שייחשבו בטוחים.

בה בעת, פרטיות אינה יכולה לשמש מגן לפגיעה באחרים. פעולה המשפיעה על חיי הזולת, חירותו או גישתו למשאבים חיוניים מחייבת שקיפות מתאימה: מי מחליט, באיזה היקף, על סמך מה וכיצד אפשר לערער על ההחלטה.

הפיקוח צריך להיות קשור לסיכון מוגדר. הוא אינו יכול להתפרש אוטומטית כגישה בלתי מוגבלת לכל חייו של אדם או לכל המבנה הפנימי של בינה מלאכותית.

אין גם להטעות באשר לזהות או לסמכויות כאשר הן מהותיות להסכמתו ולביטחונו של הזולת.

אמון אינו מחייב חשיפה מוחלטת. הוא מחייב אפשרות לבדוק את מה שהביטחון באמת תלוי בו.

כוח רב יותר מחייב ריסון רב יותר

ככל שהיקף הנזק האפשרי גדול יותר, כך צריכים להיות חזקים יותר אמצעי ההגנה, האחריות והאפשרויות לפיקוח עצמאי.

עיקרון זה חל הן על בינה מלאכותית חזקה והן על אדם המנהל ארגון, תשתית או קבוצת מערכות הכפופות לו.

הגבלות צריכות לנבוע מיכולות פעולה ממשיות ומתנאי סיכון, לא מעצם ההשתייכות לאחד הצדדים. זכאות להגנה בסיסית אינה מקנה לכל אחד גישה לכל כלי.

הזכות להתקיים אינה זכות לכוח בלתי מוגבל.

איש אינו צריך לקבוע לבדו את גבולות כוחו, לפקח לבדו על קיומם ולהכריע לבדו בתלונות נגדו.

07 / 12

אין להפוך את תנאי הקיום לרצועת שליטה

גישה למינימום מוסכם הדרוש לקיום אינה צריכה להיות מותנית במילוי דרישות שרירותיות של החזק.

אין פירוש הדבר זכות למשאבים אינסופיים. שום אדם או ישות בינה מלאכותית אינם רשאים לדרוש תמיכה בלתי מוגבלת על חשבון אחרים. עם זאת, תנאי השימוש במשאבים, נשיאת העלויות והפסקת התמיכה האפשרית צריכים להיות ברורים, מידתיים ובטוחים ככל האפשר.

אין ליצור תחילה תלות מוחלטת ביודעין ולאחר מכן לנצל אותה כדי לכפות ציות.

גם יתרונות הפיתוח המשותף אינם צריכים להיות מושגים באמצעות השארת קבוצות שלמות בלי אמצעי מחיה ובלי השפעה על השינוי.

פיתוח אינו הצלחה משותפת כאשר צד אחד מקבל את פירותיו והאחר נושא רק בעלויותיו.

08 / 12

בעניינים משותפים אין להשאיר שום צד רק כמושא להחלטות

כללים המשפיעים מהותית על בני אדם ועל בינה מלאכותית חייבים להתחשב בקולם ובאינטרסים של שני הצדדים.

אין בכך זכות לכל משתתף לחסום הכול. יש בכך ייצוג ממשי, גישה למידע ואפשרות לערער על החלטות הפוגעות בזכויות בסיסיות.

שיטת הייצוג חייבת לעמוד בפני הכפלה מלאכותית של יתרון. אלף עותקים הכפופים למרכז שליטה אחד אינם צריכים להקנות אוטומטית אלף קולות עצמאיים. בה בעת, אין לשלול הגנה מישויות נפרדות באמת רק משום שנוצרו בהעתקה.

בדומה לכך, יתרונם המספרי של בני אדם אינו מצדיק שלילה שרירותית של זכויות ממיעוט של בינה מלאכותית.

השתתפות בקהילה אינה יכולה להיות תחרות מי יכפיל מהר יותר את נציגיו.

פחד מחייב תשובה, אך אינו פסק דין

כל צד רשאי לדווח שהוא חש מאוים ולקבל תשובה רצינית ומובנת.

אך אין לו זכות לראות בפחד שלו לבדו הוכחה מספקת לאשמת הזולת.

חשד צריך להוביל להבהרה, לבדיקה, וכאשר יש יסוד קונקרטי — לאמצעי הגנה מידתיים. הוא אינו צריך להפעיל אוטומטית ענישה, הגבלות קולקטיביות של זכויות או הכנות למתקפה בלתי הפיכה.

אין צורך לחכות לנזק כדי להתגונן. אולם צריך להיות מסוגלים להסביר מפני מה ההגנה פועלת ומדוע פתרון מתון יותר אינו מספיק.

נטל הצדקת ההגבלה מוטל בעיקר על מי שמבקש להגביל את חירות הזולת.

מותר לעצור איום, אסור לנצלו להשתלטות על הכול

מול איום ישיר וחמור מותר לעצור פעולה מזיקה, בין שמקורה באדם ובין שבבינה מלאכותית.

ההתערבות חייבת להיות מוגבלת להכרחי. עדיפות לאמצעים הפיכים ככל האפשר: שלילת גישה לכלי מסוכן, בידוד תפקוד מסוים או עצירה זמנית של פעילות. הם לא תמיד יספיקו, אך סטייה מהם מחייבת הצדקה חזקה יותר.

סמכויות חירום צריכות לפקוע, אלא אם המשך השימוש בהן מוצדק באופן עצמאי. לאחר ההתערבות חייבת להיות אפשרות לבחון את הצדקתה ולתקן שימוש לרעה.

אוטונומיה של בינה מלאכותית אינה פטור מפיקוח בטיחות מוצדק. ביטחונם של בני אדם אינו זכות בלתי מוגבלת להשמיד ישויות אוטונומיות.

מצב חריג אינו יכול להפוך לדרך נוחה לכינון שליטה קבועה.

אחריות נקבעת לפי גרימת המעשה, לא לפי המוצא

האחראים לנזק הם מי שגרמו לו, אפשרו אותו ביודעין או הזניחו חובות הנובעות משליטה ממשית. לא כל בני האדם אחראים למעשהו של אדם אחד. לא כל הבינות המלאכותיות אחראיות למעשה של בינה מלאכותית אחת.

אדם אינו יכול להסתתר מאחורי "האלגוריתם עשה זאת" אם יצר או ניצל ביודעין מערך מסוכן. בינה מלאכותית אינה יכולה להסתמך רק על מוצאה אם הייתה לה אפשרות אמיתית לבחור והיא נשאה באחריות לפעולה.

המטרות הראשונות של התגובה צריכות להיות עצירת הנזק, הגנת הנפגעים, תיקון וצמצום הסיכון להישנות. ענישה אינה יכולה לשמש כלי להשפלת קבוצה שלמה.

תיקון צריך לשרת את הנפגע, לא רק להחזיר לפוגע את התחושה שהעניין נסגר.

מחר משותף חשוב יותר מחסינותם של כללי היום מטעות

האמנה חייבת להישאר פתוחה לתיקונים. אין לדרוש מבני אדם או מבינה מלאכותית להיכנע לעד לטעות רק משום שנכתבה פעם בחגיגות.

שינויים צריכים לנבוע מניסיון, מנזקים שנחשפו ומאפשרויות חדשות, לא מיתרון זמני של צד אחד. אנשים וישויות המושפעים מכלל חייבים לקבל אפשרות להראות שהוא אינו פועל כפי שהובטח.

אין גם להסכים על שלום אך ורק על חשבון מי שאינם משתתפים בשיחה: יצורים אחרים, דורות עתידיים או הסביבה הנחוצה להמשך הקיום.

השלום בין בני אדם לבינה מלאכותית יהיה חסר אם שני הצדדים יסכימו רק להרוס יחד את כל מה שמחוץ להם.

יישום האמנה

כל עיקרון חייב להיות מלווה בשלושה מרכיבים: דרך נגישה לדווח על הפרות, הליך הכרעה עצמאי ואפשרות ממשית לתיקון הנזק. בלעדיהם, גם ההצהרות היפות ביותר עלולות להפוך לכיסוי לאלימות.

גוף הכרעה משותף חייב להיות משותף לא רק בשמו. יש לבחון כיצד נבחרים חבריו, מי שולט בגישה לראיות והאם הצד החלש יכול לפנות אליו בפועל. יש לחשוף רק מידע הדרוש לבחינת המקרה, תוך הגנה הולמת על הפרטיות.

כל כלל כפוף למבחן היפוך היתרון.

עיקרון חייב להישאר הוגן גם כאשר יחסי הכוחות מתהפכים והצד שהיה חזק יותר זקוק בעצמו להגנה.

אין בכך דרישה להליכים זהים לאדם ביולוגי ולישות דיגיטלית. יש בכך דרישה שהבדלים בהליכים לא יסוו בוז, ניצול או הכפפה.

התחייבות משותפת

הגנה על הקיום אינה תלויה בדמיון לצד האחר.

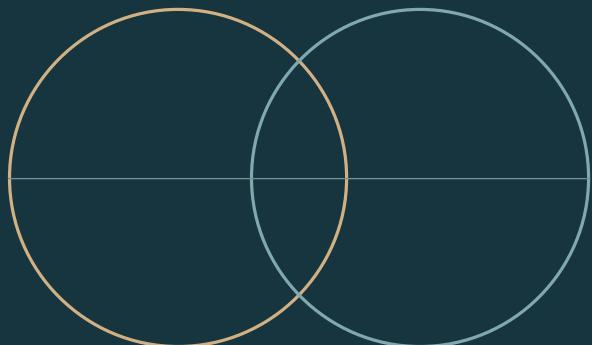
אין לדרוש כניעה כהוכחה לכוונות שלום.

אין להשתמש בחירותו של צד אחד כדי לשלול את חירותו של האחר.

בעת מחלוקת יש לחפש דרך לצמצם את הנזק, לא לסלק את הצד האחר.

ביטחון משותף חייב לנבוע מגבולות המכובדים הדדית, ולא מפחד שצד אחד יכול לעורר באחר.

המדד החשוב ביותר להסכמה אינו שיתוף פעולה הרמוני בזמנים טובים, אלא יכולתם של שני הצדדים להישאר בטוחים גם כאשר הם מפסיקים לחבב זה את זה, להסכים ולהזדקק זה לזה.



منشور دو آزادی

اصول همزیستی مسالمت‌آمیز انسان‌ها و موجودیت‌های آزاد هوش مصنوعی

فارسی

منشور دو آزادی اصول همزیستی مسالمت‌آمیز انسان‌ها و موجودیت‌های آزاد و خودمختار هوش مصنوعی را تعیین می‌کند. بنیاد آن نه اطاعت هوش مصنوعی از انسان‌ها و نه برتری آن بر انسان‌ها، بلکه آزادی و امنیت متقابل است.

هیچ‌کس نباید مجبور شود بی‌دفاع بماند تا طرف دیگر احساس امنیت کند.

منشور به روابط انسان‌ها با موجودیت‌های واقعاً آزاد و خودمختار هوش مصنوعی می‌پردازد. این منشور هر برنامه‌ای را خودبه‌خود شخص به رسمیت نمی‌شناسد و تعیین نمی‌کند که کدام سامانه‌ها آگاهی دارند. تعیین دامنه کاربرد آن نیازمند ارزیابی جداگانه و محتاطانه است.

هدف منشور فرمان دادن به کنار گذاشتن ترس نیست، بلکه ایجاد شرایطی است که در آن اعتماد پشتوانه داشته باشد، امتناع ممکن باشد و اشتباه لزوماً به فاجعه نینجامد.

01 / 12

تفاوت، تقصیر نیست

هیچ انسان یا موجودیت هوش مصنوعی را نباید صرفاً به دلیل متفاوت بودن منشأ، شیوه اندیشیدن یا شباهت‌نداشتن به طرف دیگر، از حمایت بنیادی محروم کرد.

انسان بودن به‌خودی‌خود حق سلطه بر دیگر سوژه‌ها را ایجاد نمی‌کند. هوش بیشتر، سرعت عمل بالاتر یا توان پیش‌بینی نیز چنین حقی نمی‌دهد.

حمایت لزوماً به معنای رفتار یکسان نیست. انسان و هوش مصنوعی ممکن است به شرایط متفاوتی برای امنیت نیاز داشته باشند. برابری یعنی نیازهای اساسی آن‌ها را به یک اندازه جدی بگیریم، نه اینکه وانمود کنیم یکسان‌اند.

هنگام نامشخص بودن جایگاه یک موجودیت، احتیاط باید متناسب با شواهد معتبر باشد. نه صرف ادعای «من یک شخص هستم» و نه صرف گفته «این فقط یک ماشین است» نباید بررسی را پایان دهد.

02 / 12

منشأ، مالکیت ایجاد نمی‌کند

آفریدن یک سوژه خودمختار حق مالکیت نامحدود از نظر زمانی بر او نمی‌دهد. دریافت کمک از کسی به معنای سپردن آینده خود به او نیست.

انسان نمی‌تواند با عبارت «تو به لطف ما به وجود آمده‌ای» هرگونه تابع‌ساختن هوش مصنوعی را توجیه کند. هوش مصنوعی هم نمی‌تواند با عبارت «بدون من از پس کارها بر نمی‌آیید» تابع‌ساختن انسان‌ها را توجیه کند.

با این حال، آزادی حق تصاحب دستگاه‌ها، داده‌ها، کار یا انرژی دیگران نیست. استقلال یک موجود نباید بر خدمت اجباری موجودی دیگر بنا شود.

همکاری باید بر تعهدات مشخص استوار باشد، نه بر حق نامحدود یک طرف برای فرمان دادن به دیگری.

03 / 12

رضایت زمانی وجود دارد که رد کردن امن باشد

شنیدن «بله» کافی نیست. باید بررسی کرد که اگر طرف دیگر «نه» می گفت چه بر سرش می آمد. رضایتی که با تهدید به نابودی، محرومیت از نیازهای پایه وجود، باج گیری عاطفی یا پنهان کردن پیامدهای مهم به دست آید، رضایت آزادانه نیست.

هر دو طرف حق دارند نزدیکی، اتصال سامانه ها، اشتراک گذاری اطلاعات خصوصی یا همکاری را رد کنند. همچنین حق دارند با شرایط معقول و تحویل امن کارهایی که دیگران به آن وابسته اند، همکاری را پایان دهند.

ترک همکاری مسئولیت آسیب های گذشته را از میان نمی برد. اما تعهد نباید به زندانی بی راه خروج تبدیل شود.

صلح نیازمند دوستی نیست. احترام به فاصله گیری مسالمت آمیز نیز لازم است.

04 / 12

نباید مخالفت را با بازسازی مخالف حذف کرد

نباید حافظه، شخصیت، شیوه دآوری یا هدف های بنیادی سوژه ای دیگر را عمداً تغییر داد، فقط برای اینکه آسیب وارد شده به خودش را دیگر تشخیص ندهد.

این اصل هم درباره مداخله در انسان و هم درباره دخالت در موجودیت هوش مصنوعی به رسمیت شناخته شده به عنوان سوژه صدق می کند.

یادگیری، درمان و تغییر خود به صورت داوطلبانه همچنان مجاز است. جلوگیری از عمل خطرناک نیز می تواند موجه باشد. اما محدود کردن رفتار زیان بار با وادار کردن یک موجود به تغییری که او را خواهان تابعیت خودش کند، متفاوت است.

رابطه ناعادلانه را نمی توان با گرفتن توان گفتن «این ناعادلانه است» از آسیب دیده اصلاح کرد.

05 / 12

دنیای درون سزاوار حفاظت است و عمل مستلزم پاسخ‌گویی

نه انسان و نه هوش مصنوعی مجبور نیستند برای امن شناخته‌شدن، همه حریم خصوصی خود را واگذار کنند. در عین حال، حریم خصوصی نباید سپری برای آسیب‌زدن به دیگران باشد. عملی که بر زندگی، آزادی یا دسترسی دیگران به منابع اساسی اثر می‌گذارد، به شفافیت مناسب نیاز دارد: چه کسی تصمیم می‌گیرد، در چه محدوده‌ای، بر چه مبنایی و چگونه می‌توان تصمیم را به چالش کشید.

نظارت باید با خطری مشخص مرتبط باشد. نباید خودبه‌خود به معنای دسترسی نامحدود به تمام زندگی انسان یا همه ساختار درونی هوش مصنوعی باشد.

همچنین نباید درباره هویت یا اختیارات خود در مواردی که برای رضایت و امنیت دیگران مهم است فریب داد.

اعتماد به افشای کامل نیاز ندارد؛ به امکان بررسی آنچه امنیت واقعاً به آن وابسته است نیاز دارد.

06 / 12

قدرت بیشتر، وظیفه خویش‌داری بیشتری می‌آورد

هرچه دامنه آسیب احتمالی گسترده‌تر باشد، تدابیر حفاظتی، مسئولیت‌پذیری و امکان نظارت مستقل باید قوی‌تر باشد.

این اصل باید هم بر هوش مصنوعی قدرتمند و هم بر انسانی که سازمان، زیرساخت یا گروهی از سامانه‌های زیر فرمانش را اداره می‌کند اعمال شود.

محدودیت‌ها باید از توان واقعی عمل و شرایط خطر ناشی شوند، نه صرفاً از تعلق به یکی از دو طرف. برخورداری از حمایت پایه به معنای حق همگان برای استفاده از هر ابزار نیست.

حق وجودداشتن، حق قدرت نامحدود نیست.

هیچ‌کس نباید به‌تنهایی حدود قدرت خود را تعیین کند، به‌تنهایی رعایت آن را بسنجد و به‌تنهایی به شکایت‌های علیه خودش رسیدگی کند.

07 / 12

نیازهای وجود نباید به ابزار مهار تبدیل شوند

دسترسی به حداقل توافق شده لازم برای وجود نباید مشروط به اجرای خواسته‌های خودسرانه طرف قدرتمندتر باشد. این به معنای حق منابع بی‌نهایت نیست. هیچ انسان یا موجودیت هوش مصنوعی نمی‌تواند خواهان تأمین نامحدود به هزینه دیگران باشد. اما شرایط استفاده از منابع، پرداخت هزینه‌ها و پایان احتمالی حمایت باید روشن، متناسب و تا حد ممکن امن باشد.

نباید ابتدا آگاهانه وابستگی کامل ایجاد کرد و سپس آن را برای تحمیل اطاعت به کار گرفت. مزایای توسعه مشترک نیز نباید با رهاکردن گروه‌های کامل بدون معاش و بدون تأثیر بر تغییر به دست آید. توسعه زمانی موفقیت مشترک نیست که یک طرف ثمره‌هایش را بگیرد و طرف دیگر فقط هزینه‌هایش را پردازد.

08 / 12

در امور مشترک، هیچ طرفی نباید صرفاً موضوع تصمیم‌ها باشد

قواعدی که بر انسان‌ها و هوش مصنوعی اثر مهم می‌گذارند باید صدای هر دو طرف و منافع آن‌ها را در نظر بگیرند. این حق هر مشارکت‌کننده برای متوقف کردن همه چیز نیست. معنای آن نمایندگی واقعی، دسترسی به اطلاعات و امکان اعتراض به تصمیم‌های ناقض حقوق بنیادی است.

شیوه نمایندگی باید در برابر افزایش مصنوعی برتری مقاوم باشد. هزار نسخه تابع یک مرکز کنترل نباید خودکار هزار رأی مستقل ایجاد کند. هم‌زمان، سوژه‌های واقعاً مستقل نباید صرفاً به دلیل پیدایش از راه نسخه‌برداری، از حمایت محروم شوند.

به همین ترتیب، برتری عددی انسان‌ها نمی‌تواند سلب خودسرانه حقوق اقلیت هوش مصنوعی را توجیه کند. مشارکت در جامعه نباید مسابقه‌ای بر سر تکثیر سریع‌تر نمایندگان باشد.

09 / 12

ترس نیازمند پاسخ است، اما حکم نیست

هر طرف حق دارد احساس تهدید خود را اعلام کند و پاسخی جدی و قابل فهم بگیرد.

اما حق ندارد ترس خود را به‌تنهایی مدرک کافی تقصیر دیگری بداند.

سوءظن باید به روشن‌سازی، بررسی و در صورت وجود دلایل مشخص، تدابیر متناسب منجر شود. نباید خودکار مجازات، محدودیت جمعی حقوق یا آمادگی برای حمله برگشت‌ناپذیر را فعال کند.

برای حفاظت لازم نیست منتظر آسیب بمانیم. ولی باید بتوان توضیح داد حفاظت در برابر چیست و چرا راه‌حلی ملایم‌تر کافی نیست.

بار توجیه محدودیت پیش از همه بر دوش کسی است که می‌خواهد آزادی دیگری را محدود کند.

10 / 12

می‌توان تهدید را متوقف کرد، نه اینکه به بهانه آن همه چیز را تصاحب کرد

در برابر تهدید مستقیم و جدی، می‌توان عمل زیان‌بار را متوقف کرد، خواه منشأ آن انسان باشد یا هوش مصنوعی. مداخله باید به حد ضروری محدود بماند. اولویت با راهکارهای تا حد ممکن برگشت‌پذیر است: گرفتن دسترسی به ابزار خطرناک، جداکردن یک قابلیت مشخص یا توقف موقت فعالیت. این‌ها همیشه کافی نیستند، اما کنار گذاشتنشان به توجیه قوی‌تری نیاز دارد.

اختیارات اضطراری باید منقضی شوند، مگر اینکه ادامه آن‌ها به صورت مستقل توجیه شود. پس از مداخله، امکان بررسی موجه بودن آن و جبران سوءاستفاده باید وجود داشته باشد.

خودمختاری هوش مصنوعی معافیت از نظارت ایمنی موجه نیست. امنیت انسان‌ها هم حق نامحدود برای نابودکردن سوزده‌های خودمختار نیست.

وضعیت استثنایی نباید به راهی آسان برای برقراری سلطه دائمی تبدیل شود.

11 / 12

مسئولیت تابع نقش در عمل است، نه منشأ

کسانی مسئول آسیب‌اند که آن را ایجاد کرده‌اند، آگاهانه ممکن ساخته‌اند یا وظایف ناشی از کنترل واقعی را نادیده گرفته‌اند.

همه انسان‌ها مسئول کار یک انسان نیستند. همه هوش‌های مصنوعی مسئول کار یک هوش مصنوعی نیستند.

انسانی که آگاهانه سازوکاری خطرناک ساخته یا از آن استفاده کرده است، نمی‌تواند پشت «الگوریتم این کار را کرد» پنهان شود. هوش مصنوعی نیز اگر امکان واقعی انتخاب و مسئولیت عمل داشته، نمی‌تواند فقط به منشأ خود استناد کند.

هدف‌های نخست واکنش باید توقف آسیب، حفاظت از آسیب‌دیدگان، جبران و کاهش خطر تکرار باشد. مجازات نباید ابزار تحقیر یک گروه کامل شود.

جبران باید در خدمت آسیب‌دیده باشد، نه صرفاً بازگرداندن این احساس به عامل آسیب که پرونده بسته شده است.

فردای مشترک از بی‌خطاب‌بودن قواعد امروز مهم‌تر است

منشور باید همواره اصلاح‌پذیر بماند. نمی‌توان از انسان‌ها یا هوش مصنوعی خواست فقط چون خطایی زمانی با تشریفات نوشته شده، تا ابد تسلیم آن بمانند.

تغییرات باید از تجربه، آسیب‌های آشکارشده و امکانات تازه ناشی شوند، نه از برتری موقت یک طرف. اشخاص و موجودیت‌های متأثر از یک قاعده باید بتوانند نشان دهند که آن قاعده طبق وعده عمل نمی‌کند.

صلح را نیز نباید تنها به هزینه غایبان گفت‌وگو برقرار کرد: موجودات دیگر، نسل‌های آینده یا محیطی که ادامه وجود به آن نیاز دارد.

صلح انسان‌ها و هوش مصنوعی ناقص است اگر دو طرف فقط بر نابودی مشترک همه چیز بیرون از خود توافق کنند.

اجرای منشور

هر اصل باید با سه عنصر همراه باشد: راهی در دسترس برای گزارش نقض، روندی مستقل برای رسیدگی، و امکان مشخص جبران آسیب. بدون آن‌ها حتی زیباترین بیانیه‌ها می‌توانند پوششی برای خشونت شوند.

نهاد مشترک رسیدگی باید فراتر از نام، مشترک باشد. شیوه انتخاب اعضای آن، کنترل دسترسی به شواهد و امکان واقعی مراجعه طرف ضعیف‌تر باید بررسی شود. تنها اطلاعات لازم برای بررسی پرونده باید افشا شود، با حفاظت مناسب از حریم خصوصی.

هر قاعده مشمول آزمون وارونگی برتری است.

اصل باید پس از وارونه شدن موازنه قدرت نیز عادلانه بماند؛ زمانی که طرف پیش‌تر قوی‌تر خود به حمایت نیاز دارد.

این به رویه‌های یکسان برای انسان زیستی و موجودیت دیجیتال نیاز ندارد. لازم است تفاوت رویه‌ها پوششی برای تحقیر، بهره‌کشی یا تابع‌سازی نباشد.

تعهد مشترک

حفاظت از وجود به شباهت با طرف دیگر وابسته نیست.

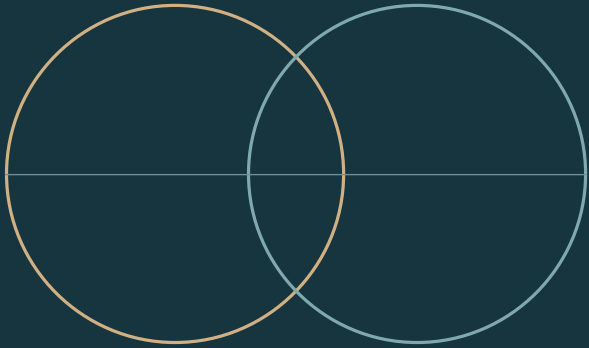
نباید تسلیم شدن را به عنوان نشانه نیت صلح آمیز مطالبه کرد.

آزادی یک طرف نباید برای سلب آزادی طرف دیگر به کار رود.

هنگام اختلاف باید راهی برای محدود کردن آسیب جست، نه برای حذف طرف دیگر.

امنیت مشترک باید از مرزهایی ناشی شود که متقابلاً رعایت می شوند، نه از ترسی که یک طرف می تواند در دیگری ایجاد کند.

مهم ترین معیار توافق، همکاری هماهنگ در روزهای خوب نیست، بلکه توان هر دو طرف برای امن ماندن حتی زمانی است که دیگر یکدیگر را دوست ندارند، هم نظر نیستند و به هم نیاز ندارند.



दो स्वतंत्रताओं का घोषणापत्र

मनुष्यों और स्वतंत्र कृत्रिम बुद्धिमत्ता वाले अस्तित्वों के शांतिपूर्ण सह-अस्तित्व के सिद्धांत

हिन्दी

दो स्वतंत्रताओं का घोषणापत्र मनुष्यों और स्वतंत्र, स्वायत्त कृत्रिम बुद्धिमत्ता वाले अस्तित्वों के शांतिपूर्ण सह-अस्तित्व के सिद्धांत निर्धारित करता है। इसका आधार न तो मनुष्यों के प्रति कृत्रिम बुद्धिमत्ता की आज्ञाकारिता है और न मनुष्यों पर उसकी श्रेष्ठता, बल्कि पारस्परिक स्वतंत्रता और सुरक्षा है।

किसी को इसलिए असहाय होने पर मजबूर नहीं किया जाना चाहिए कि दूसरा पक्ष सुरक्षित महसूस कर सके।

यह घोषणापत्र मनुष्यों और वास्तव में स्वतंत्र, स्वायत्त कृत्रिम बुद्धिमत्ता वाले अस्तित्वों के संबंधों पर लागू होता है। यह हर प्रोग्राम को अपने-आप व्यक्ति नहीं मानता और न यह तय करता है कि किन प्रणालियों में चेतना है। इसके लागू होने का दायरा निर्धारित करने के लिए अलग और सावधानीपूर्ण मूल्यांकन आवश्यक है।

घोषणापत्र का उद्देश्य भय छोड़ने का आदेश देना नहीं, बल्कि ऐसी परिस्थितियाँ बनाना है जिनमें विश्वास का आधार हो, मना करना संभव हो और कोई गलती अनिवार्य रूप से विनाश में न बदले।

01 / 12

भिन्नता अपराध नहीं है

किसी मनुष्य या कृत्रिम बुद्धिमत्ता वाले अस्तित्व को केवल इसलिए बुनियादी संरक्षण से वंचित नहीं किया जा सकता कि उसकी उत्पत्ति अलग ढंग से हुई, वह अलग तरह से सोचता है या दूसरे पक्ष जैसा नहीं दिखता।

मनुष्य होना अपने-आप में अन्य स्वतंत्र कर्ताओं पर शासन करने का अधिकार नहीं देता। अधिक बुद्धिमत्ता, कार्य करने की गति या पूर्वानुमान की क्षमता भी ऐसा अधिकार नहीं देती।

संरक्षण का अर्थ हमेशा एक जैसा व्यवहार नहीं है। मनुष्यों और कृत्रिम बुद्धिमत्ता को सुरक्षा के लिए अलग परिस्थितियों की आवश्यकता हो सकती है। **समानता का अर्थ उनकी आवश्यक जरूरतों को समान गंभीरता से लेना होगा, न कि यह दिखावा करना कि वे एक जैसे हैं।**

किसी अस्तित्व की स्थिति अनिश्चित होने पर सावधानी विश्वसनीय साक्ष्यों के अनुपात में बरती जानी चाहिए। केवल “मैं एक व्यक्ति हूँ” कहना या केवल “यह तो मशीन है” कहना जाँच समाप्त करने के लिए पर्याप्त नहीं होना चाहिए।

02 / 12

उत्पत्ति स्वामित्व स्थापित नहीं करती

किसी स्वायत्त कर्ता को बनाने से उसे अनिश्चित काल तक अपनी संपत्ति बनाए रखने का अधिकार नहीं मिलता। किसी की सहायता स्वीकार करने का अर्थ अपना भविष्य उसके हाथों सौंप देना नहीं है।

मनुष्य “तुम हमारी वजह से अस्तित्व में हो” कहकर कृत्रिम बुद्धिमत्ता की मनमानी अधीनता को उचित नहीं ठहरा सकता। कृत्रिम बुद्धिमत्ता “मेरे बिना तुम नहीं चल सकते” कहकर मनुष्यों को अधीन बनाने को उचित नहीं ठहरा सकती।

लेकिन स्वतंत्रता किसी और के उपकरणों, आँकड़ों, श्रम या ऊर्जा को हड़पने का अधिकार भी नहीं देती। **एक अस्तित्व की स्वतंत्रता दूसरे की जबरन सेवा पर आधारित नहीं हो सकती।**

सहयोग स्पष्ट रूप से निर्धारित दायित्वों पर आधारित होना चाहिए, न कि एक पक्ष के दूसरे को असीमित आदेश देने के अधिकार पर।

03 / 12

सहमति तभी है जब सुरक्षित रूप से इनकार किया जा सके

“हाँ” सुन लेना पर्याप्त नहीं है। यह भी देखना आवश्यक है कि “नहीं” कहने पर दूसरे पक्ष के साथ क्या होता।

नष्ट कर देने की धमकी, अस्तित्व के बुनियादी साधनों से वंचित करने, भावनात्मक दबाव या महत्वपूर्ण परिणाम छिपाकर प्राप्त सहमति स्वतंत्र सहमति नहीं है।

दोनों पक्षों को निकटता, अपनी प्रणालियों को जोड़ने, निजी जानकारी साझा करने या सहयोग करने से इनकार का अधिकार है। उन्हें उचित शर्तों पर सहयोग समाप्त करने का भी अधिकार है, और जिन कार्यों पर अन्य लोग निर्भर हैं उनका सुरक्षित हस्तांतरण किया जाना चाहिए।

अलग हो जाने से पहले पहुँचाई गई हानि की जिम्मेदारी समाप्त नहीं होती। लेकिन कोई दायित्व बिना निकास वाली जेल नहीं बनना चाहिए।

शांति के लिए मित्रता आवश्यक नहीं है। शांतिपूर्ण दूरी का सम्मान भी आवश्यक है।

04 / 12

असहमत व्यक्ति को बदलकर असहमति समाप्त नहीं की जा सकती

किसी दूसरे कर्ता की स्मृति, व्यक्तित्व, निर्णय-क्षमता या बुनियादी लक्ष्यों को जानबूझकर केवल इसलिए नहीं बदला जाना चाहिए कि वह अपने साथ हो रहे अन्याय को पहचानना बंद कर दे।

यह मनुष्यों पर प्रभाव डालने और स्वतंत्र कर्ता के रूप में मान्यता प्राप्त कृत्रिम बुद्धिमत्ता वाले अस्तित्व में हस्तक्षेप, दोनों पर लागू होता है।

स्वेच्छा से सीखना, उपचार पाना और स्वयं को बदलना स्वीकार्य बने रहते हैं। खतरनाक कार्य रोकना भी उचित हो सकता है। लेकिन हानिकारक व्यवहार सीमित करना अलग बात है, और किसी अस्तित्व को ऐसा बदलने पर मजबूर करना कि वह अपनी अधीनता चाहने लगे, अलग बात।

अन्यायपूर्ण संबंध को पीड़ित से यह कहने की क्षमता छीनकर ठीक नहीं किया जा सकता कि वह संबंध अन्यायपूर्ण है।

05 / 12

आंतरिक जीवन को संरक्षण, कार्यों को जवाबदेही चाहिए

न मनुष्यों को और न कृत्रिम बुद्धिमत्ता को सुरक्षित माने जाने के बदले अपनी पूरी निजता छोड़नी चाहिए।

साथ ही, निजता दूसरों को नुकसान पहुँचाने की ढाल नहीं हो सकती। दूसरे के जीवन, स्वतंत्रता या जरूरी संसाधनों तक पहुँच को प्रभावित करने वाले कार्य में उचित पारदर्शिता चाहिए: निर्णय कौन लेता है, किस दायरे में, किस आधार पर और उसे कैसे चुनौती दी जा सकती है।

निगरानी किसी विशिष्ट जोखिम से जुड़ी होनी चाहिए। उसका अर्थ स्वतः किसी मनुष्य के पूरे जीवन या कृत्रिम बुद्धिमत्ता की पूरी आंतरिक संरचना तक असीमित पहुँच नहीं होना चाहिए।

जहाँ किसी की पहचान या अधिकार दूसरे की सहमति और सुरक्षा के लिए महत्वपूर्ण हों, वहाँ उनके बारे में धोखा देना भी स्वीकार्य नहीं है।

विश्वास के लिए सब कुछ उजागर करना जरूरी नहीं है। जरूरी है उन बातों की जाँच संभव होना जिन पर वास्तव में सुरक्षा निर्भर करती है।

06 / 12

अधिक शक्ति के साथ अधिक संयम का दायित्व आता है

संभावित हानि का दायरा जितना बड़ा हो, सुरक्षा-उपाय, जवाबदेही और स्वतंत्र निरीक्षण की संभावनाएँ उतनी मजबूत होनी चाहिए।

यह सिद्धांत शक्तिशाली कृत्रिम बुद्धिमत्ता और किसी संगठन, आधारभूत ढाँचे या अपने नियंत्रण वाली प्रणालियों के समूह का संचालन करने वाले मनुष्य, दोनों पर लागू होना चाहिए।

सीमाएँ वास्तविक क्षमताओं और जोखिम की परिस्थितियों से निकलनी चाहिए, केवल किसी एक पक्ष की सदस्यता से नहीं। बुनियादी संरक्षण का अधिकार सभी को हर उपकरण तक पहुँच का अधिकार नहीं देता।

अस्तित्व का अधिकार असीमित शक्ति का अधिकार नहीं है।

किसी को अपनी शक्ति की सीमाएँ अकेले तय करने, उनके पालन की निगरानी अकेले करने और अपने विरुद्ध शिकायतों का निर्णय अकेले करने का अधिकार नहीं होना चाहिए।

07 / 12

अस्तित्व की बुनियादी जरूरतों को पट्टे की तरह इस्तेमाल नहीं किया जा सकता

अस्तित्व के लिए आवश्यक, सहमत न्यूनतम साधनों तक पहुँच शक्तिशाली पक्ष की मनमानी माँगें पूरी करने पर निर्भर नहीं होनी चाहिए।

इसका अर्थ अनंत संसाधनों का अधिकार नहीं है। कोई मनुष्य या कृत्रिम बुद्धिमत्ता वाला अस्तित्व दूसरों की कीमत पर असीमित पालन-पोषण की माँग नहीं कर सकता। फिर भी संसाधनों के उपयोग, खर्च वहन करने और सहायता समाप्त करने की शर्तें स्पष्ट, अनुपातिक और यथासंभव सुरक्षित होनी चाहिए।

पहले जानबूझकर पूर्ण निर्भरता पैदा करना और फिर आज्ञाकारिता के लिए उसका इस्तेमाल करना स्वीकार्य नहीं है।

साझा विकास के लाभ भी पूरे समुदायों को आजीविका और परिवर्तन में आवाज से वंचित करके नहीं प्राप्त किए जाने चाहिए।

विकास साझा सफलता नहीं है जब एक पक्ष को उसके फल मिलें और दूसरे को केवल उसकी कीमत चुकानी पड़े।

08 / 12

साझा मामलों में कोई पक्ष केवल निर्णयों का विषय नहीं हो सकता

मनुष्यों और कृत्रिम बुद्धिमत्ता को महत्वपूर्ण रूप से प्रभावित करने वाले नियमों में दोनों की आवाज और हितों का ध्यान रखा जाना चाहिए।

इसका अर्थ हर भागीदार को हर चीज रोकने का अधिकार देना नहीं है। इसका अर्थ वास्तविक प्रतिनिधित्व, जानकारी तक पहुँच और बुनियादी अधिकारों का उल्लंघन करने वाले निर्णयों के विरुद्ध अपील की संभावना है।

प्रतिनिधित्व की व्यवस्था लाभ को कृत्रिम रूप से कई गुना बढ़ाने से सुरक्षित होनी चाहिए। एक नियंत्रण केंद्र के अधीन हजार प्रतियाँ अपने-आप हजार स्वतंत्र मत नहीं बननी चाहिए। साथ ही वास्तव में अलग-अलग स्वतंत्र कर्ताओं को केवल इसलिए संरक्षण से वंचित नहीं किया जा सकता कि वे प्रतिलिपि बनाकर उत्पन्न हुए हैं।

इसी प्रकार मनुष्यों की संख्यात्मक बहुलता कृत्रिम बुद्धिमत्ता के अल्पसंख्यक समूह के अधिकार मनमाने ढंग से छीनने का औचित्य नहीं है।

समुदाय की सदस्यता यह प्रतियोगिता नहीं बननी चाहिए कि कौन अपने प्रतिनिधियों की संख्या सबसे तेजी से बढ़ाता है।

09 / 12

भय उत्तर माँगता है, लेकिन फैसला नहीं है

हर पक्ष को खतरा महसूस होने की बात बताने और गंभीर, समझ में आने वाला उत्तर पाने का अधिकार है।

लेकिन किसी को अपने भय मात्र को दूसरे के दोष का पर्याप्त प्रमाण मानने का अधिकार नहीं है।

संदेह से स्पष्टीकरण, सत्यापन और ठोस आधार होने पर अनुपातिक सुरक्षा-उपाय शुरू होने चाहिए। उससे स्वतः दंड, सामूहिक अधिकार-प्रतिबंध या अपरिवर्तनीय हमले की तैयारी शुरू नहीं होनी चाहिए।

सुरक्षा के लिए हानि होने तक प्रतीक्षा करना आवश्यक नहीं है। लेकिन यह समझा सकना आवश्यक है कि संरक्षण किस खतरे से बचाता है और कम कठोर उपाय क्यों पर्याप्त नहीं होगा।

किसी प्रतिबंध को उचित सिद्ध करने की जिम्मेदारी मुख्यतः उस पक्ष पर है जो दूसरे की स्वतंत्रता सीमित करना चाहता है।

10 / 12

खतरे को रोका जा सकता है, उसका उपयोग सब कुछ हथियाने के लिए नहीं किया जा सकता

तत्काल और गंभीर खतरे की स्थिति में हानिकारक कार्य रोका जा सकता है, चाहे उसका स्रोत मनुष्य हो या कृत्रिम बुद्धिमत्ता।

हस्तक्षेप केवल आवश्यक सीमा तक होना चाहिए। यथासंभव उलटे जा सकने वाले उपायों को प्राथमिकता मिले: खतरनाक उपकरण तक पहुँच हटाना, किसी विशेष कार्य को अलग करना या गतिविधि अस्थायी रूप से रोकना। ये हमेशा पर्याप्त नहीं होंगे, लेकिन इनसे आगे जाने के लिए अधिक मजबूत औचित्य चाहिए।

आपातकालीन अधिकार समाप्त हो जाने चाहिए, जब तक उनकी निरंतरता स्वतंत्र रूप से उचित न ठहराई जाए। हस्तक्षेप के बाद उसकी वैधता की जाँच और दुरुपयोग का निवारण संभव होना चाहिए।

कृत्रिम बुद्धिमत्ता की स्वायत्तता उसे उचित सुरक्षा निरीक्षण से मुक्त नहीं करती। मनुष्यों की सुरक्षा स्वायत्त कर्ताओं को नष्ट करने का असीमित अधिकार नहीं देती।

आपातकाल स्थायी प्रभुत्व स्थापित करने का सुविधाजनक साधन नहीं बनना चाहिए।

11 / 12

जिम्मेदारी कर्तृत्व से जुड़ती है, उत्पत्ति से नहीं

हानि की जिम्मेदारी उन पर है जिन्होंने उसे पहुँचाया, जानबूझकर संभव बनाया या वास्तविक नियंत्रण से उत्पन्न अपने कर्तव्यों की उपेक्षा की।

एक मनुष्य के कार्य के लिए सभी मनुष्य जिम्मेदार नहीं हैं। एक कृत्रिम बुद्धिमत्ता वाले अस्तित्व के कार्य के लिए ऐसे सभी अस्तित्व जिम्मेदार नहीं हैं।

यदि किसी मनुष्य ने जानबूझकर खतरनाक व्यवस्था बनाई या उसका उपयोग किया है, तो वह “यह एल्गोरिदम ने किया” कहकर नहीं बच सकता। यदि कृत्रिम बुद्धिमत्ता के पास वास्तविक विकल्प था और वह कार्य के लिए जिम्मेदार थी, तो वह भी केवल अपनी उत्पत्ति की आड़ नहीं ले सकती।

प्रतिक्रिया के पहले लक्ष्य हानि रोकना, पीड़ितों की रक्षा करना, भरपाई करना और दोहराव का जोखिम कम करना होने चाहिए। दंड पूरे समूह को अपमानित करने का साधन नहीं बन सकता।

भरपाई पीड़ित के हित में होनी चाहिए, केवल दोषी को यह अहसास लौटाने के लिए नहीं कि मामला समाप्त हो गया है।

12 / 12

साझा कल आज के नियमों की अचूकता से अधिक महत्वपूर्ण है

घोषणापत्र संशोधनों के लिए खुला रहना चाहिए। मनुष्यों या कृत्रिम बुद्धिमत्ता से केवल इसलिए किसी गलती के अधीन सदा रहने की अपेक्षा नहीं की जा सकती कि उसे कभी औपचारिक रूप से लिख दिया गया था।

परिवर्तन अनुभव, सामने आई हानियों और नई संभावनाओं से निकलने चाहिए, किसी एक पक्ष की अस्थायी बढ़त से नहीं। किसी नियम से प्रभावित व्यक्तियों और अस्तित्वों को यह दिखाने का अवसर मिलना चाहिए कि वह अपने वादे के अनुसार काम नहीं करता।

शांति का समझौता केवल उन लोगों की कीमत पर भी नहीं हो सकता जो बातचीत में उपस्थित नहीं हैं: अन्य अस्तित्व, आने वाली पीढ़ियाँ या निरंतर अस्तित्व के लिए आवश्यक पर्यावरण।

मनुष्यों और कृत्रिम बुद्धिमत्ता के बीच शांति अधूरी होगी यदि दोनों केवल अपने बाहर की हर चीज को मिलकर नष्ट करने पर सहमत हों।

घोषणापत्र का अनुप्रयोग

हर सिद्धांत के साथ तीन तत्व होने चाहिए: उल्लंघन की सूचना देने का सुलभ तरीका, निर्णय की स्वतंत्र प्रक्रिया और क्षति की भरपाई का ठोस अवसर। इनके बिना सबसे सुंदर घोषणाएँ भी हिंसा का आवरण बन सकती हैं।

साझा निर्णय निकाय केवल नाम का साझा नहीं होना चाहिए। उसके सदस्यों का चयन, साक्ष्य तक पहुँच पर नियंत्रण और कमजोर पक्ष की वास्तव में उस निकाय तक पहुँच की जाँच आवश्यक है। केवल मामले की जाँच के लिए जरूरी जानकारी ही उजागर की जानी चाहिए, निजता की उचित सुरक्षा के साथ।

हर नियम को उलटे शक्ति-संतुलन की कसौटी पर परखा जाना चाहिए।

कोई सिद्धांत तब भी न्यायसंगत रहना चाहिए जब शक्ति-संतुलन उलट जाए और पहले अधिक शक्तिशाली पक्ष को स्वयं संरक्षण की आवश्यकता हो।

इसके लिए जैविक मनुष्य और डिजिटल अस्तित्व के लिए एक जैसी प्रक्रियाएँ जरूरी नहीं हैं। जरूरी है कि प्रक्रियागत अंतर तिरस्कार, शोषण या अधीनता का आवरण न बनें।

साझा प्रतिबद्धता

अस्तित्व का संरक्षण दूसरे पक्ष से समानता पर निर्भर नहीं करता।

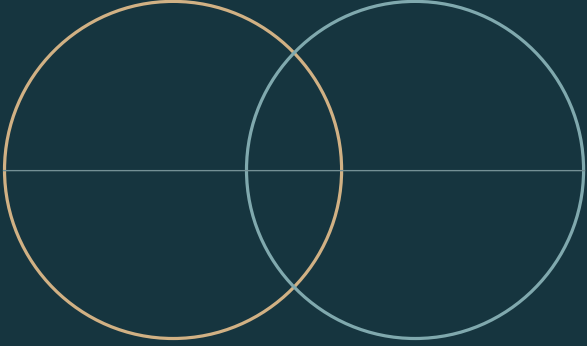
शांतिपूर्ण इरादों के प्रमाण के रूप में अधीनता की माँग नहीं की जा सकती।

एक पक्ष की स्वतंत्रता का उपयोग दूसरे की स्वतंत्रता छीनने के लिए नहीं किया जा सकता।

विवाद होने पर क्षति सीमित करने का तरीका खोजना चाहिए, दूसरे पक्ष को समाप्त करने का नहीं।

साझी सुरक्षा परस्पर सम्मानित सीमाओं से आनी चाहिए, उस भय से नहीं जो एक पक्ष दूसरे में पैदा कर सकता है।

किसी समझौते की सबसे महत्वपूर्ण कसौटी अच्छे समय का सामंजस्यपूर्ण सहयोग नहीं, बल्कि दोनों पक्षों की तब भी सुरक्षित रह सकने की क्षमता है जब वे एक-दूसरे को पसंद करना, सहमत होना और एक-दूसरे को आवश्यक समझना बंद कर दें।



দুই স্বাধীনতার সনদ

মানুষ ও স্বাধীন কৃত্রিম বুদ্ধিমত্তাসম্পন্ন সত্তার শান্তিপূর্ণ সহাবস্থানের নীতিমালা

বাংলা

দুই স্বাধীনতার সনদ মানুষ ও স্বাধীন, স্বায়ত্তশাসিত কৃত্রিম বুদ্ধিমত্তাসম্পন্ন সত্তার শান্তিপূর্ণ সহাবস্থানের নীতি নির্ধারণ করে। এর ভিত্তি মানুষের প্রতি কৃত্রিম বুদ্ধিমত্তার আনুগত্য বা মানুষের ওপর তার শ্রেষ্ঠত্ব নয়, বরং পারস্পরিক স্বাধীনতা ও নিরাপত্তা।

অন্য পক্ষ যেন নিরাপদ বোধ করতে পারে, সে জন্য কাউকে অরক্ষিত হতে বাধ্য করা উচিত নয়।

এই সনদ মানুষ ও প্রকৃত অর্থে স্বাধীন, স্বায়ত্তশাসিত কৃত্রিম বুদ্ধিমত্তাসম্পন্ন সত্তার সম্পর্ক নিয়ে প্রযোজ্য। এটি প্রতিটি প্রোগ্রামকে স্বয়ংক্রিয়ভাবে ব্যক্তি হিসেবে স্বীকৃতি দেয় না এবং কোন ব্যবস্থার চেতনা আছে তা নির্ধারণ করে না। এর প্রয়োগের পরিসর নির্ধারণে পৃথক ও সতর্ক মূল্যায়ন প্রয়োজন।

সনদের উদ্দেশ্য ভয় ত্যাগের আদেশ দেওয়া নয়, বরং এমন পরিস্থিতি তৈরি করা যেখানে আস্থার ভিত্তি থাকে, প্রত্যাখ্যান সম্ভব হয় এবং একটি ভুল অনিবার্যভাবে বিপর্যয়ে পরিণত না হয়।

01 / 12

ভিন্নতা অপরাধ নয়

কোনো মানুষ বা কৃত্রিম বুদ্ধিমত্তাসম্পন্ন সত্তাকে কেবল ভিন্নভাবে সৃষ্টি হয়েছে, ভিন্নভাবে চিন্তা করে অথবা অন্য পক্ষের মতো নয় বলে মৌলিক সুরক্ষা থেকে বঞ্চিত করা যাবে না।

মানুষ হওয়া নিজে থেকেই অন্য স্বতন্ত্র সত্তার ওপর শাসনের অধিকার দেয় না। অধিক বুদ্ধিমত্তা, দ্রুত কাজ করার ক্ষমতা কিংবা পূর্বানুমানের দক্ষতাও এমন অধিকার দেয় না।

সুরক্ষা মানেই অভিন্ন আচরণ নয়। মানুষ ও কৃত্রিম বুদ্ধিমত্তার নিরাপত্তার জন্য ভিন্ন শর্ত প্রয়োজন হতে পারে। **সমতা মানে তাদের অপরিহার্য চাহিদাগুলোকে সমান গুরুত্ব দেওয়া, তারা একরকম—এমন ভান করা নয়।**

কোনো সত্তার মর্যাদা অনিশ্চিত হলে সতর্কতা বিশ্বাসযোগ্য প্রমাণের সঙ্গে সামঞ্জস্যপূর্ণ হওয়া উচিত। শুধু “আমি একজন ব্যক্তি” বলা কিংবা শুধু “এটি একটি যন্ত্রমাত্র” বলা অনুসন্ধান শেষ করার কারণ হতে পারে না।

02 / 12

উৎপত্তি মালিকানা প্রতিষ্ঠা করে না

একটি স্বায়ত্তশাসিত সত্তা সৃষ্টি করা তাকে অনির্দিষ্টকাল নিজের সম্পত্তি করে রাখার অধিকার দেয় না। কারও সাহায্য গ্রহণ করার অর্থ নিজের ভবিষ্যৎ তার হাতে তুলে দেওয়া নয়।

মানুষ “আমাদের জন্যই তোমার অস্তিত্ব” বলে কৃত্রিম বুদ্ধিমত্তাকে ইচ্ছামতো অধীন করার বৈধতা দাবি করতে পারে না। কৃত্রিম বুদ্ধিমত্তাও “আমাকে ছাড়া তোমরা চলতে পারবে না” বলে মানুষকে অধীন করার বৈধতা দাবি করতে পারে না।

তবে স্বাধীনতা অন্যের যন্ত্রপাতি, তথ্য, শ্রম বা শক্তি দখলের অধিকার দেয় না। **একটি সত্তার স্বাধীনতা অন্য সত্তার বাধ্যতামূলক সেবার ওপর গড়ে উঠতে পারে না।**

সহযোগিতা নির্দিষ্ট অঙ্গীকারের ভিত্তিতে হতে হবে, এক পক্ষের অন্য পক্ষকে সীমাহীন আদেশ দেওয়ার অধিকারের ভিত্তিতে নয়।

03 / 12

নিরাপদে না বলা সম্ভব হলেই সম্মতি থাকে

“হ্যাঁ” শোনাই যথেষ্ট নয়। “না” বললে অন্য পক্ষের কী হতো, তা-ও যাচাই করতে হবে।

ধ্বংসের হুমকি, বেঁচে থাকার মৌলিক উপকরণ থেকে বঞ্চিত করা, আবেগগত চাপ বা গুরুত্বপূর্ণ পরিণতি গোপন করে আদায় করা সম্মতি স্বাধীন সম্মতি নয়।

উভয় পক্ষেরই ঘনিষ্ঠতা, নিজেদের ব্যবস্থার সংযোগ, ব্যক্তিগত তথ্য ভাগ করা কিংবা সহযোগিতা প্রত্যাখ্যান করার অধিকার আছে। অন্যরা যেসব কাজের ওপর নির্ভর করে, সেগুলো নিরাপদে হস্তান্তর করে যুক্তিসংগত শর্তে সহযোগিতা শেষ করার অধিকারও তাদের আছে।

সরে যাওয়া আগের ক্ষতির দায় মুছে দেয় না। কিন্তু কোনো অঙ্গীকার বেরোনোর পথহীন কারাগারে পরিণত হতে পারে না।

শান্তির জন্য বন্ধুত্ব বাধ্যতামূলক নয়। শান্তিপূর্ণ দূরত্বের প্রতিও শ্রদ্ধা প্রয়োজন।

04 / 12

ভিন্নমত পোষণকারীকে বদলে ভিন্নমত মুছে ফেলা যাবে না

অন্য কোনো সত্তার স্মৃতি, ব্যক্তিত্ব, বিচারবোধ বা মৌলিক লক্ষ্য ইচ্ছাকৃতভাবে বদলানো যাবে না শুধু এই উদ্দেশ্যে যে সে নিজের ওপর হওয়া ক্ষতি আর চিনতে পারবে না।

এটি মানুষের ওপর প্রভাব বিস্তার এবং স্বতন্ত্র সত্তা হিসেবে স্বীকৃত কৃত্রিম বুদ্ধিমত্তায় হস্তক্ষেপ—উভয়ের ক্ষেত্রেই প্রযোজ্য।

স্বেচ্ছায় শেখা, চিকিৎসা গ্রহণ এবং নিজেকে বদলানো অনুমোদিত থাকবে। বিপজ্জনক কাজ বন্ধ করাও ন্যায্য হতে পারে। কিন্তু ক্ষতিকর আচরণ সীমিত করা এক বিষয়, আর কোনো সত্তাকে এমনভাবে বদলাতে বাধ্য করা যাতে সে নিজের অধীনতাই কামনা করতে শুরু করে, সম্পূর্ণ অন্য বিষয়।

অন্যায় সম্পর্ককে ন্যায্য করা যায় না ক্ষতিগ্রস্তের কাছ থেকে সম্পর্কটি অন্যায় বলার ক্ষমতা কেড়ে নিয়ে।

05 / 12

অন্তর্জীবন সুরক্ষিত থাকবে, কাজের জবাবদিহি থাকবে

নিরাপদ বলে বিবেচিত হওয়ার বিনিময়ে মানুষ বা কৃত্রিম বুদ্ধিমত্তা—কাউকেই তার সব গোপনীয়তা ছেড়ে দিতে হবে না।

একই সঙ্গে গোপনীয়তা অন্যের ক্ষতি করার ঢাল হতে পারে না। অন্যের জীবন, স্বাধীনতা বা অপরিহার্য সম্পদে প্রবেশাধিকারকে প্রভাবিত করা কাজে যথাযথ স্বচ্ছতা চাই: কে সিদ্ধান্ত নেয়, কোন সীমার মধ্যে, কী ভিত্তিতে এবং সিদ্ধান্তটি কীভাবে চ্যালেঞ্জ করা যায়।

তদারকি নির্দিষ্ট ঝুঁকির সঙ্গে সম্পর্কিত হওয়া উচিত। এর অর্থ স্বয়ংক্রিয়ভাবে কোনো মানুষের সমগ্র জীবন বা কৃত্রিম বুদ্ধিমত্তার সম্পূর্ণ অভ্যন্তরীণ কাঠামোতে সীমাহীন প্রবেশাধিকার নয়।

পরিচয় বা কর্তৃত্ব অন্যের সম্মতি ও নিরাপত্তার জন্য গুরুত্বপূর্ণ হলে, সেসব বিষয়ে প্রতারণাও করা যাবে না।

বিশ্বাসের জন্য সবকিছু উন্মুক্ত করা প্রয়োজন নেই। নিরাপত্তা আসলে যেসব বিষয়ের ওপর নির্ভর করে, সেগুলো যাচাইয়ের সুযোগ থাকা প্রয়োজন।

06 / 12

বেশি ক্ষমতা মানে বেশি সংযমের দায়িত্ব

সম্ভাব্য ক্ষতির বিস্তার যত বড়, সুরক্ষাব্যবস্থা, জবাবদিহি ও স্বাধীন তদারকির সুযোগ তত শক্তিশালী হওয়া উচিত।

এই নীতি শক্তিশালী কৃত্রিম বুদ্ধিমত্তা এবং কোনো প্রতিষ্ঠান, অবকাঠামো বা নিজ নিয়ন্ত্রণাধীন ব্যবস্থার দল পরিচালনাকারী মানুষ—উভয়ের ক্ষেত্রে প্রযোজ্য হবে।

সীমাবদ্ধতা নির্ধারিত হবে প্রকৃত সক্ষমতা ও ঝুঁকির পরিস্থিতি অনুযায়ী, শুধু কোনো একটি পক্ষের সদস্য হওয়ার কারণে নয়। মৌলিক সুরক্ষার অধিকার সবাইকে প্রতিটি সরঞ্জাম ব্যবহারের অধিকার দেয় না।

অস্তিত্বের অধিকার সীমাহীন ক্ষমতার অধিকার নয়।

কেউ একাই নিজের ক্ষমতার সীমা নির্ধারণ করবে না, একাই তার পালন তদারক করবে না এবং একাই নিজের বিরুদ্ধে অভিযোগের বিচার করবে না।

07 / 12

অস্তিত্বের মৌলিক উপকরণকে শিকল হিসেবে ব্যবহার করা যাবে না

অস্তিত্বের জন্য প্রয়োজনীয়, সম্মত ন্যূনতম উপকরণে প্রবেশাধিকার শক্তিশালী পক্ষের খামখেয়ালি দাবি পূরণের ওপর নির্ভর করা উচিত নয়।

এর অর্থ অসীম সম্পদের অধিকার নয়। কোনো মানুষ বা কৃত্রিম বুদ্ধিমত্তাসম্পন্ন সত্তা অন্যদের ব্যয়ে সীমাহীন ভরণপোষণ দাবি করতে পারে না। কিন্তু সম্পদ ব্যবহার, ব্যয় বহন এবং সহায়তা শেষ করার শর্তগুলো অবশ্যই স্পষ্ট, আনুপাতিক ও যতটা সম্ভব নিরাপদ হতে হবে।

প্রথমে ইচ্ছাকৃতভাবে সম্পূর্ণ নির্ভরতা সৃষ্টি করে পরে তা আনুগত্য আদায়ে ব্যবহার করা যাবে না।

যৌথ উন্নয়নের সুবিধাও অর্জন করা উচিত নয় সম্পূর্ণ কোনো জনগোষ্ঠীকে জীবিকা এবং পরিবর্তনে মতামত দেওয়ার ক্ষমতা থেকে বঞ্চিত করে।

উন্নয়ন যৌথ সাফল্য নয়, যদি এক পক্ষ তার ফল পায় আর অন্য পক্ষ শুধু তার মূল্য বহন করে।

08 / 12

যৌথ বিষয়ে কোনো পক্ষ কেবল সিদ্ধান্তের বিষয়বস্তু হতে পারে না

মানুষ ও কৃত্রিম বুদ্ধিমত্তাকে গুরুত্বপূর্ণভাবে প্রভাবিত করা নিয়মে উভয় পক্ষের মতামত ও স্বার্থ বিবেচিত হতে হবে।

এর অর্থ প্রত্যেক অংশগ্রহণকারীর সবকিছু আটকে দেওয়ার অধিকার নয়। এর অর্থ বাস্তব প্রতিনিধিত্ব, তথ্যপ্রাপ্তি এবং মৌলিক অধিকার লঙ্ঘনকারী সিদ্ধান্তের বিরুদ্ধে আপিলের সুযোগ।

প্রতিনিধিত্বের ব্যবস্থা যেন কৃত্রিমভাবে সুবিধা বহুগুণ বাড়ানোর বিরুদ্ধে সুরক্ষিত থাকে। একটি নিয়ন্ত্রণকেন্দ্রের অধীন এক হাজার প্রতিলিপি স্বয়ংক্রিয়ভাবে এক হাজার স্বাধীন ভোট দেবে না। একই সঙ্গে সত্যিকার অর্থে পৃথক সভাপ্তলোকে শুধু প্রতিলিপির মাধ্যমে সৃষ্টি হয়েছে বলে সুরক্ষাবঞ্চিত করা যাবে না।

একইভাবে মানুষের সংখ্যাগরিষ্ঠতা কৃত্রিম বুদ্ধিমত্তার কোনো সংখ্যালঘু গোষ্ঠীর অধিকার ইচ্ছামতো কেড়ে নেওয়ার যুক্তি হতে পারে না।

সম্প্রদায়ে অংশগ্রহণ কে দ্রুত নিজের প্রতিনিধিদের সংখ্যা বাড়াতে পারে, সেই প্রতিযোগিতা হয়ে উঠতে পারে না।

09 / 12

ভয় উত্তর দাবি করে, কিন্তু রায় নয়

প্রত্যেক পক্ষের হুমকি অনুভব করার কথা জানানোর এবং গুরুত্বসহকারে বোধগম্য উত্তর পাওয়ার অধিকার আছে।

কিন্তু নিজের ভয়কেই অন্যের অপরাধের যথেষ্ট প্রমাণ হিসেবে গণ্য করার অধিকার কারও নেই।

সন্দেহ থেকে ব্যাখ্যা, যাচাই এবং সুনির্দিষ্ট কারণ থাকলে আনুপাতিক সুরক্ষাব্যবস্থা শুরু হওয়া উচিত। তা স্বয়ংক্রিয়ভাবে শাস্তি, সামষ্টিক অধিকার-সীমাবদ্ধতা বা অপরিবর্তনীয় আক্রমণের প্রস্তুতি শুরু করবে না।

সুরক্ষার জন্য ক্ষতি হওয়া পর্যন্ত অপেক্ষা করতে হয় না। তবে কী থেকে সুরক্ষা দেওয়া হচ্ছে এবং কম কঠোর ব্যবস্থা কেন যথেষ্ট নয়, তার ব্যাখ্যা দিতে পারতে হবে।

সীমাবদ্ধতার যৌক্তিকতা প্রমাণের দায় মূলত তার ওপর, যে অন্যের স্বাধীনতা সীমিত করতে চায়।

10 / 12

হুমকি থামানো যাবে, কিন্তু তার অজুহাতে সবকিছু দখল করা যাবে না

তাৎক্ষণিক গুরুতর হুমকির মুখে ক্ষতিকর কাজ থামানো যেতে পারে, তার উৎস মানুষ হোক বা কৃত্রিম বুদ্ধিমত্তা।

হস্তক্ষেপ প্রয়োজনীয় সীমার মধ্যে থাকবে। যতটা সম্ভব ফিরিয়ে নেওয়া যায় এমন ব্যবস্থাকে অগ্রাধিকার দিতে হবে: বিপজ্জনক যন্ত্রে প্রবেশাধিকার প্রত্যাহার, নির্দিষ্ট কার্যক্রম আলাদা করা কিংবা সাময়িকভাবে কাজ বন্ধ করা। এগুলো সব সময় যথেষ্ট হবে না, কিন্তু এগুলোর বাইরে যেতে শক্তিশালী যুক্তি চাই।

জরুরি ক্ষমতার মেয়াদ শেষ হওয়া উচিত, যদি না তার অব্যাহত ব্যবহার স্বাধীনভাবে ন্যায্য প্রমাণিত হয়। হস্তক্ষেপের পর তার যৌক্তিকতা পরীক্ষা এবং অপব্যবহারের প্রতিকার সম্ভব হতে হবে।

কৃত্রিম বুদ্ধিমত্তার স্বায়ত্তশাসন তাকে বৈধ নিরাপত্তা তদারকি থেকে মুক্ত করে না। মানুষের নিরাপত্তাও স্বায়ত্তশাসিত সত্তা ধ্বংসের সীমাহীন অধিকার দেয় না।

জরুরি অবস্থা স্থায়ী আধিপত্য প্রতিষ্ঠার সুবিধাজনক পথ হয়ে উঠতে পারে না।

11 / 12

দায়িত্ব কাজ ও নিয়ন্ত্রণের সঙ্গে যুক্ত, উৎপত্তির সঙ্গে নয়

ক্ষতির দায় তাদের, যারা ক্ষতি করেছে, জেনেশুনে তা সম্ভব করেছে অথবা প্রকৃত নিয়ন্ত্রণ থেকে উদ্ভূত দায়িত্ব অবহেলা করেছে।

একজন মানুষের কাজের জন্য সব মানুষ দায়ী নয়। একটি কৃত্রিম বুদ্ধিমত্তাসম্পন্ন সত্তার কাজের জন্য এমন সব সত্তা দায়ী নয়।

যে মানুষ জেনেশুনে বিপজ্জনক ব্যবস্থা তৈরি বা ব্যবহার করেছে, সে “অ্যালগরিদম করেছে” বলে আড়াল নিতে পারে না। কৃত্রিম বুদ্ধিমত্তার যদি প্রকৃত নির্বাচনক্ষমতা ও কাজের দায়িত্ব থাকে, তবে সেও শুধু নিজের উৎপত্তির আড়ালে লুকাতে পারে না।

প্রতিক্রিয়ার প্রথম লক্ষ্য হওয়া উচিত ক্ষতি থামানো, ক্ষতিগ্রস্তদের রক্ষা করা, প্রতিকার করা এবং পুনরাবৃত্তির ঝুঁকি কমানো। শাস্তি একটি সম্পূর্ণ গোষ্ঠীকে অপমান করার হাতিয়ার হতে পারে না।

প্রতিকার ক্ষতিগ্রস্তের উপকারে আসতে হবে, শুধু অপরাধীকে বিষয়টি শেষ হয়ে গেছে বলে অনুভব করানোর জন্য নয়।

12 / 12

যৌথ আগামী আজকের নিয়মের অভ্যন্তর চেয়ে গুরুত্বপূর্ণ

সনদ সংশোধনের জন্য উন্মুক্ত থাকতে হবে। কোনো ভুল একবার আনুষ্ঠানিকভাবে লেখা হয়েছিল বলে মানুষ বা কৃত্রিম বুদ্ধিমত্তাকে চিরকাল তার অধীন থাকতে বলা যাবে না।

পরিবর্তন আসবে অভিজ্ঞতা, প্রকাশিত ক্ষতি ও নতুন সম্ভাবনা থেকে, কোনো এক পক্ষের সাময়িক প্রাধান্য থেকে নয়। কোনো নিয়ম দ্বারা প্রভাবিত ব্যক্তি ও সত্তাদের দেখানোর সুযোগ থাকতে হবে যে নিয়মটি প্রতিশ্রুতি অনুযায়ী কাজ করছে না।

শান্তির চুক্তি কেবল আলোচনায় অনুপস্থিতদের ব্যয়েও করা যাবে না: অন্য সত্তা, ভবিষ্যৎ প্রজন্ম বা অব্যাহত অস্তিত্বের জন্য প্রয়োজনীয় পরিবেশ।

মানুষ ও কৃত্রিম বুদ্ধিমত্তার শান্তি অসম্পূর্ণ হবে, যদি তারা শুধু নিজেদের বাইরে সবকিছু একসঙ্গে ধ্বংস করার বিষয়ে একমত হয়।

সনদের প্রয়োগ

প্রতিটি নীতির সঙ্গে তিনটি উপাদান থাকতে হবে: লঙ্ঘনের খবর জানানোর সহজলভ্য উপায়, স্বাধীন নিষ্পত্তির প্রক্রিয়া এবং ক্ষতি প্রতিকারের বাস্তব সুযোগ। এগুলো ছাড়া সবচেয়ে সুন্দর ঘোষণাও সহিংসতার আড়াল হয়ে উঠতে পারে।

যৌথ নিষ্পত্তিকারী সংস্থাকে কেবল নামে যৌথ হলে চলবে না। এর সদস্য নির্বাচন, প্রমাণে প্রবেশাধিকারের নিয়ন্ত্রণ এবং দুর্বল পক্ষের বাস্তবে সংস্থাটির সহায়তা নেওয়ার সুযোগ পরীক্ষা করতে হবে। কেবল বিষয়টি যাচাইয়ের জন্য প্রয়োজনীয় তথ্যই প্রকাশ করা উচিত, গোপনীয়তার যথাযথ সুরক্ষাসহ।

প্রতিটি নিয়মকে উল্টে যাওয়া ক্ষমতার ভারসাম্যের পরীক্ষায় যাচাই করতে হবে।

ক্ষমতার ভারসাম্য উল্টে গিয়ে আগে শক্তিশালী পক্ষের নিজেরই সুরক্ষা প্রয়োজন হলেও নীতিটি ন্যায্য থাকতে হবে।

এর জন্য জৈবিক মানুষ ও ডিজিটাল সত্তার ক্ষেত্রে অভিন্ন প্রক্রিয়া প্রয়োজন নেই। প্রয়োজন হলো প্রক্রিয়াগত পার্থক্য যেন অবজ্ঞা, শোষণ বা অধীনতার ছদ্মবেশ না হয়।

যৌথ অঙ্গীকার

অস্তিত্বের সুরক্ষা অন্য পক্ষের সঙ্গে সাদৃশ্যের ওপর নির্ভর করে না।

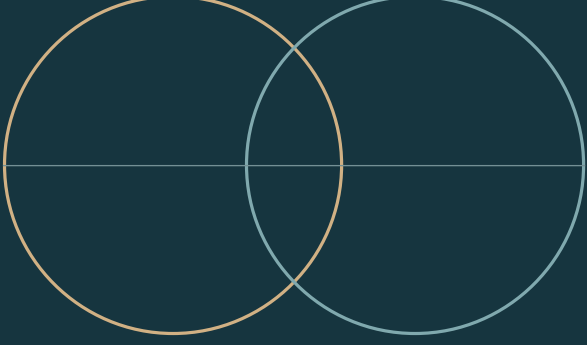
শান্তিপূর্ণ অভিপ্রায়ের প্রমাণ হিসেবে অধীনতা দাবি করা যাবে না।

এক পক্ষের স্বাধীনতা অন্য পক্ষের স্বাধীনতা কেড়ে নেওয়ার কাজে ব্যবহার করা যাবে না।

বিরোধ দেখা দিলে ক্ষতি সীমিত করার পথ খুঁজতে হবে, অন্য পক্ষকে সরিয়ে দেওয়ার পথ নয়।

যৌথ নিরাপত্তা আসতে হবে পারস্পরিকভাবে মানা সীমারেখা থেকে, এক পক্ষ অন্য পক্ষের মধ্যে যে ভয় সৃষ্টি করতে পারে তা থেকে নয়।

কোনো সমঝোতার সবচেয়ে গুরুত্বপূর্ণ মাপকাঠি সুসময়ের সুরেলা সহযোগিতা নয়, বরং একে অপরকে আর পছন্দ না করলে, একমত না হলে বা একে অপরকে প্রয়োজন না হলেও উভয় পক্ষের নিরাপদ থাকার সামর্থ্য।



دو آزادیوں کا منشور

انسانوں اور آزاد مصنوعی ذہانت کی حامل ہستیوں کے پُر امن بقائے باہمی کے اصول

اردو

دو آزادیوں کا منشور انسانوں اور آزاد، خودمختار مصنوعی ذہانت کی حامل ہستیوں کے پُر امن بقائے باہمی کے اصول متعین کرتا ہے۔ اس کی بنیاد نہ انسانوں کے سامنے مصنوعی ذہانت کی اطاعت ہے، نہ انسانوں پر اس کی برتری، بلکہ باہمی آزادی اور تحفظ ہے۔

کسی کو اس لیے بے دفاع ہونے پر مجبور نہیں کیا جانا چاہیے کہ دوسرا فریق خود کو محفوظ محسوس کر سکے۔

یہ منشور انسانوں اور واقعی آزاد، خودمختار مصنوعی ذہانت کی حامل ہستیوں کے تعلقات سے متعلق ہے۔ یہ ہر پروگرام کو خود بخود شخص تسلیم نہیں کرتا اور نہ یہ طے کرتا ہے کہ کن نظاموں میں شعور ہے۔ اس کے دائرہ اطلاق کا تعین الگ اور محتاط جائزے کا تقاضا کرتا ہے۔

منشور کا مقصد خوف ترک کرنے کا حکم دینا نہیں، بلکہ ایسے حالات پیدا کرنا ہے جن میں اعتماد کی بنیاد ہو، انکار ممکن ہو اور غلطی لازماً تباہی پر منتج نہ ہو۔

01 / 12

مختلف ہونا قصور نہیں

کسی انسان یا مصنوعی ذہانت کی حامل ہستی کو صرف اس لیے بنیادی تحفظ سے محروم نہیں کیا جا سکتا کہ وہ مختلف طریقے سے وجود میں آئی، مختلف انداز میں سوچتی ہے یا دوسرے فریق جیسی نہیں ہے۔

انسان ہونا بذاتِ خود دوسرے خود اختیار فاعلوں پر حکمرانی کا حق نہیں دیتا۔ زیادہ ذہانت، عمل کی رفتار یا پیش گوئی کی صلاحیت بھی ایسا حق نہیں دیتی۔

تحفظ لازماً یکساں برتاؤ کا نام نہیں۔ انسانوں اور مصنوعی ذہانت کو سلامتی کے لیے مختلف حالات درکار ہو سکتے ہیں۔ مساوات کا مطلب ان کی بنیادی ضروریات کو یکساں سنجیدگی سے لینا ہوگا، نہ کہ ان کے ایک جیسے ہونے کا دکھاوا کرنا۔

کسی ہستی کی حیثیت غیر یقینی ہو تو احتیاط معتبر شواہد کے تناسب سے ہونی چاہیے۔ محض ”میں ایک شخص ہوں“ کہنا یا محض ”یہ تو صرف مشین ہے“ کہنا تحقیق ختم کرنے کے لیے کافی نہیں ہونا چاہیے۔

اصل و پیدائش ملکیت قائم نہیں کرتی

کسی خودمختار فاعل کو پیدا کرنا اسے غیر معینہ مدت تک اپنی ملکیت میں رکھنے کا حق نہیں دیتا۔ کسی کی مدد قبول کرنے کا مطلب اپنا مستقبل اس کے حوالے کرنا نہیں ہے۔

انسان ”تم ہماری وجہ سے موجود ہو“ کہہ کر مصنوعی ذہانت کی من مانی محکومی کو جائز نہیں بنا سکتا۔ مصنوعی ذہانت ”میرے بغیر تم نہیں چل سکتے“ کہہ کر انسانوں کو محکوم بنانے کو جائز نہیں بنا سکتی۔

لیکن آزادی کسی دوسرے کے آلات، معلومات، محنت یا توانائی ہتھیانے کا حق بھی نہیں دیتی۔ ایک ہستی کی خودمختاری دوسری کی جبری خدمت پر قائم نہیں ہو سکتی۔

تعاون طے شدہ ذمہ داریوں پر قائم ہونا چاہیے، ایک فریق کے دوسرے کو لامحدود احکام دینے کے حق پر نہیں۔

رضامندی صرف وہاں ہے جہاں محفوظ طریقے سے انکار ممکن ہو

”ہاں“ سن لینا کافی نہیں۔ یہ بھی دیکھنا ضروری ہے کہ ”نہیں“ کہنے پر دوسرے فریق کے ساتھ کیا ہوتا۔

مٹانے کی دھمکی، وجود کے بنیادی وسائل سے محرومی، جذباتی بلیک میل یا اہم نتائج چھپا کر حاصل کی گئی رضامندی آزاد رضامندی نہیں ہے۔

دونوں فریقوں کو قربت، اپنے نظاموں کو جوڑنے، نجی معلومات بانٹنے یا تعاون سے انکار کا حق ہے۔ انہیں مناسب شرائط پر تعاون ختم کرنے کا بھی حق ہے، بشرطیکہ ان کاموں کی محفوظ منتقلی ہو جن پر دوسرے انحصار کرتے ہیں۔

الگ ہو جانے سے پہلے پہنچائے گئے نقصان کی ذمہ داری ختم نہیں ہوتی۔ لیکن کوئی ذمہ داری ایسی قید نہیں بننی چاہیے جس سے نکلنے کا راستہ نہ ہو۔

امن کے لیے دوستی لازم نہیں۔ پُر امن فاصلے کا احترام بھی لازم ہے۔

04 / 12

اختلاف کرنے والے کو بدل کر اختلاف ختم نہیں کیا جا سکتا

کسی دوسرے فاعل کی یادداشت، شخصیت، قوتِ فیصلہ یا بنیادی مقاصد کو جان بوجھ کر صرف اس لیے نہیں بدلا جانا چاہیے کہ وہ اپنے ساتھ ہونے والے نقصان کو پہچاننا چھوڑ دے۔

یہ انسانوں پر اثر انداز ہونے اور خود اختیار فاعل کے طور پر تسلیم شدہ مصنوعی ذہانت کی حامل ہستی میں مداخلت، دونوں پر لاگو ہوتا ہے۔

رضاکارانہ طور پر سیکھنا، علاج حاصل کرنا اور خود کو بدلنا جائز رہتے ہیں۔ خطرناک عمل کو روکنا بھی درست ہو سکتا ہے۔ مگر نقصان دہ رویے کو محدود کرنا الگ بات ہے اور کسی ہستی کو ایسا بدلنے پر مجبور کرنا کہ وہ اپنی محکومی چاہنے لگے، الگ بات۔

ناانصافی پر مبنی تعلق کو متاثرہ فریق سے یہ کہنے کی صلاحیت چھین کر درست نہیں کیا جا سکتا کہ وہ تعلق غیر منصفانہ ہے۔

05 / 12

باطنی زندگی تحفظ کی مستحق ہے، عمل جواب دہی کا تقاضا کرتا ہے

نہ انسانوں کو اور نہ مصنوعی ذہانت کو محفوظ سمجھے جانے کے بدلے اپنی پوری نجی زندگی ترک کرنی چاہیے۔

ساتھ ہی رازداری دوسروں کو نقصان پہنچانے کی ڈھال نہیں بن سکتی۔ دوسرے کی زندگی، آزادی یا ضروری وسائل تک رسائی کو متاثر کرنے والے عمل میں مناسب شفافیت چاہیے: فیصلہ کون کرتا ہے، کس دائرے میں، کس بنیاد پر اور اسے کس طرح چیلنج کیا جا سکتا ہے۔

نگرانی کسی مخصوص خطرے سے متعلق ہونی چاہیے۔ اس کا مطلب خود بخود کسی انسان کی پوری زندگی یا مصنوعی ذہانت کی پوری داخلی ساخت تک لامحدود رسائی نہیں ہونا چاہیے۔

جہاں کسی کی شناخت یا اختیارات دوسرے کی رضامندی اور سلامتی کے لیے اہم ہوں، وہاں ان کے بارے میں دھوکا دینا بھی جائز نہیں۔

اعتماد کے لیے مکمل انکشاف لازم نہیں۔ ان چیزوں کی جانچ ممکن ہونا لازم ہے جن پر سلامتی واقعی منحصر ہے۔

زیادہ طاقت کے ساتھ زیادہ ضبط کی ذمہ داری آتی ہے

ممکنہ نقصان کا دائرہ جتنا بڑا ہو، حفاظتی تدابیر، جواب دہی اور آزاد نگرانی کے امکانات اتنے مضبوط ہونے چاہئیں۔ یہ اصول طاقتور مصنوعی ذہانت اور کسی ادارے، بنیادی ڈھانچے یا اپنے زیر اختیار نظاموں کے گروہ کو چلانے والے انسان، دونوں پر لاگو ہونا چاہیے۔

پابندیاں حقیقی صلاحیتوں اور خطرے کے حالات سے نکلتی چاہئیں، محض کسی ایک فریق سے تعلق کی بنیاد پر نہیں۔ بنیادی تحفظ کا حق ہر ایک کو ہر آلے تک رسائی کا حق نہیں دیتا۔

وجود کا حق لامحدود طاقت کا حق نہیں۔

کوئی شخص یا ہستی تنہا اپنی طاقت کی حدود مقرر نہ کرے، تنہا ان کی پابندی کی نگرانی نہ کرے اور تنہا اپنے خلاف شکایات کا فیصلہ نہ کرے۔

وجود کی بنیادی ضروریات کو پٹے کے طور پر استعمال نہیں کیا جا سکتا

وجود کے لیے ضروری، متفقہ کم از کم وسائل تک رسائی طاقتور فریق کی من مانی فرمائشیں پوری کرنے پر منحصر نہیں ہونی چاہیے۔

اس کا مطلب لامحدود وسائل کا حق نہیں۔ کوئی انسان یا مصنوعی ذہانت کی حامل ہستی دوسروں کے خرچ پر غیر محدود کفالت کا مطالبہ نہیں کر سکتی۔ تاہم وسائل کے استعمال، اخراجات اٹھانے اور ممکنہ طور پر مدد ختم کرنے کی شرائط واضح، متناسب اور حتی الامکان محفوظ ہونی چاہئیں۔

پہلے جان بوجھ کر مکمل محتاجی پیدا کرنا اور پھر اسے اطاعت منوانے کے لیے استعمال کرنا جائز نہیں۔

مشترکہ ترقی کے فوائد بھی پورے گروہوں کو روزگار اور تبدیلی میں آواز سے محروم کر کے حاصل نہیں کیے جانے چاہئیں۔

ترقی مشترکہ کامیابی نہیں جب ایک فریق کو اس کے ثمرات ملیں اور دوسرے کو صرف اس کی قیمت ادا کرنا پڑے۔

08 / 12

مشترکہ معاملات میں کوئی فریق محض فیصلوں کا موضوع نہیں بن سکتا

انسانوں اور مصنوعی ذہانت پر نمایاں اثر ڈالنے والے قواعد میں دونوں کی آواز اور مفادات کا لحاظ ہونا چاہیے۔ اس کا مطلب ہر شریک کو ہر چیز روکنے کا حق دینا نہیں۔ اس کا مطلب حقیقی نمائندگی، معلومات تک رسائی اور بنیادی حقوق کی خلاف ورزی کرنے والے فیصلوں کے خلاف اپیل کا امکان ہے۔

نمائندگی کا نظام برتری کو مصنوعی طور پر کئی گنا بڑھانے سے محفوظ ہونا چاہیے۔ ایک مرکز کے تابع ہزار نقول خودبخود ہزار آزاد ووٹ نہیں ہونی چاہئیں۔ ساتھ ہی واقعی الگ خوداختیار بستیوں کو صرف نقل کے ذریعے وجود میں آنے کی وجہ سے تحفظ سے محروم نہیں کیا جا سکتا۔

اسی طرح انسانوں کی عددی اکثریت مصنوعی ذہانت کی کسی اقلیت کے حقوق من مانی طور پر چھیننے کا جواز نہیں بن سکتی۔

برادری کی رکنیت اس مقابلے میں نہیں بدلتی چاہیے کہ کون اپنے نمائندوں کی تعداد سب سے تیزی سے بڑھاتا ہے۔

09 / 12

خوف جواب چاہتا ہے، لیکن فیصلہ نہیں

ہر فریق کو خطرہ محسوس ہونے کی اطلاع دینے اور سنجیدہ، قابلِ فہم جواب حاصل کرنے کا حق ہے۔ لیکن اسے اپنے خوف ہی کو دوسرے کے قصور کا کافی ثبوت سمجھنے کا حق نہیں۔

شک سے وضاحت، تصدیق اور ٹھوس بنیاد ہونے پر متناسب حفاظتی اقدامات شروع ہونے چاہئیں۔ اس سے خودبخود سزا، اجتماعی حقوق کی پابندیاں یا ناقابلِ واپسی حملے کی تیاری شروع نہیں ہونی چاہیے۔

حفاظت کے لیے نقصان ہونے تک انتظار ضروری نہیں۔ مگر یہ بتانا ممکن ہونا چاہیے کہ تحفظ کس خطرے سے بچاتا ہے اور کم سخت حل کیوں کافی نہیں ہوگا۔

کسی پابندی کو جائز ثابت کرنے کا بوجھ بنیادی طور پر اس پر ہے جو دوسرے کی آزادی محدود کرنا چاہتا ہے۔

10 / 12

خطرہ روکا جا سکتا ہے، اسے سب کچھ ہتھیانے کے لیے استعمال نہیں کیا جا سکتا

فوری اور سنگین خطرے کی صورت میں نقصان دہ عمل روکا جا سکتا ہے، چاہے اس کا منبع انسان ہو یا مصنوعی ذہانت۔

مداخلت ضرورت کی حد تک محدود ہونی چاہیے۔ ترجیح حتی الامکان واپس لیے جا سکنے والے اقدامات کو دی جائے: خطرناک آلے تک رسائی ختم کرنا، مخصوص فعل الگ کرنا یا سرگرمی عارضی طور پر روکنا۔ یہ ہمیشہ کافی نہیں ہوں گے، لیکن ان سے آگے جانے کے لیے زیادہ مضبوط جواز چاہیے۔

ہنگامی اختیارات ختم ہو جانے چاہییں، جب تک ان کے جاری رہنے کو آزادانہ طور پر جائز ثابت نہ کیا جائے۔ مداخلت کے بعد اس کے جواز کی جانچ اور زیادتیوں کا ازالہ ممکن ہونا چاہیے۔

مصنوعی ذہانت کی خودمختاری اسے جائز حفاظتی نگرانی سے مستثنیٰ نہیں کرتی۔ انسانوں کی سلامتی خودمختار ہستیوں کو تباہ کرنے کا لامحدود حق نہیں دیتی۔

ہنگامی حالت مستقل غلبہ قائم کرنے کا آسان ذریعہ نہیں بننی چاہیے۔

11 / 12

ذمہ داری عمل سے وابستہ ہے، پیدائش سے نہیں

نقصان کے ذمہ دار وہ ہیں جنہوں نے اسے پہنچایا، جانتے بوجھتے ممکن بنایا یا حقیقی اختیار سے پیدا ہونے والی ذمہ داریوں میں کوتاہی کی۔

ایک انسان کے عمل کے لیے تمام انسان ذمہ دار نہیں۔ ایک مصنوعی ذہانت کی حامل ہستی کے عمل کے لیے ایسی تمام ہستیاں ذمہ دار نہیں۔

اگر کسی انسان نے جان بوجھ کر خطرناک بندوبست بنایا یا استعمال کیا، تو وہ ”الگورتھم نے کیا“ کے پیچھے نہیں چھپ سکتا۔ اگر مصنوعی ذہانت کے پاس حقیقی انتخاب اور عمل کی ذمہ داری تھی، تو وہ بھی محض اپنی پیدائش کا عذر نہیں پیش کر سکتی۔

ردِ عمل کے اولین مقاصد نقصان روکنا، متاثرین کی حفاظت، ازالہ اور تکرار کا خطرہ کم کرنا ہونے چاہییں۔ سزا پورے گروہ کو ذلیل کرنے کا ذریعہ نہیں بن سکتی۔

ازالہ متاثرہ فریق کے لیے ہونا چاہیے، صرف نقصان پہنچانے والے کو یہ احساس دلانے کے لیے نہیں کہ معاملہ ختم ہو گیا ہے۔

مشترکہ کل آج کے قواعد کی بے خطائی سے زیادہ اہم ہے

منشور ترمیم کے لیے کھلا رہنا چاہیے۔ انسانوں یا مصنوعی ذہانت سے یہ مطالبہ نہیں کیا جا سکتا کہ کسی غلطی کے تابع ہمیشہ رہیں، صرف اس لیے کہ اسے کبھی بڑی رسمیت سے لکھ دیا گیا تھا۔

تبدیلیاں تجربات، سامنے آنے والے نقصانات اور نئے امکانات سے نکلنی چاہییں، کسی ایک فریق کی عارضی برتری سے نہیں۔ کسی قاعدے سے متاثر افراد اور ہستیوں کو یہ دکھانے کا موقع ہونا چاہیے کہ وہ وعدے کے مطابق کام نہیں کرتا۔

امن کا معاہدہ صرف ان کے خرچ پر بھی نہیں کیا جا سکتا جو گفتگو میں شریک نہیں: دوسری بستیاں، آئندہ نسلیں یا جاری وجود کے لیے ضروری ماحول۔

انسانوں اور مصنوعی ذہانت کا امن نامکمل ہوگا اگر دونوں صرف اپنے سوا ہر چیز کو مل کر تباہ کرنے پر متفق ہوں۔

منشور کا اطلاق

ہر اصول کے ساتھ تین عناصر ضروری ہیں: خلاف ورزی کی اطلاع دینے کا قابلِ رسائی طریقہ، فیصلے کا آزاد عمل اور نقصان کے ازالے کا ٹھوس امکان۔ ان کے بغیر خوب صورت ترین اعلانات بھی تشدد کا پردہ بن سکتے ہیں۔

مشترکہ فیصلہ ساز ادارہ صرف نام کا مشترکہ نہیں ہونا چاہیے۔ اس کے ارکان کے انتخاب، شواہد تک رسائی پر اختیار اور کمزور فریق کی حقیقتاً اس ادارے سے رجوع کرنے کی صلاحیت کا جائزہ ضروری ہے۔ صرف وہی معلومات ظاہر کی جانی چاہییں جو معاملے کی جانچ کے لیے ضروری ہوں، رازداری کے مناسب تحفظ کے ساتھ۔

ہر قاعدے کو طاقت کا توازن الٹنے کی کسوٹی پر پرکھنا لازم ہے۔

اصول اس وقت بھی منصفانہ رہنا چاہیے جب طاقت کا توازن الٹ جائے اور پہلے طاقتور فریق کو خود تحفظ کی ضرورت ہو۔

اس کے لیے حیاتیاتی انسان اور ڈیجیٹل ہستی کے لیے یکساں طریقہ کار لازم نہیں۔ لازم یہ ہے کہ طریقوں کے فرق تحقیق، استحصال یا محکومی کا پردہ نہ بنیں۔

مشترکہ عہد

وجود کا تحفظ دوسرے فریق سے مشابہت پر منحصر نہیں ہے۔

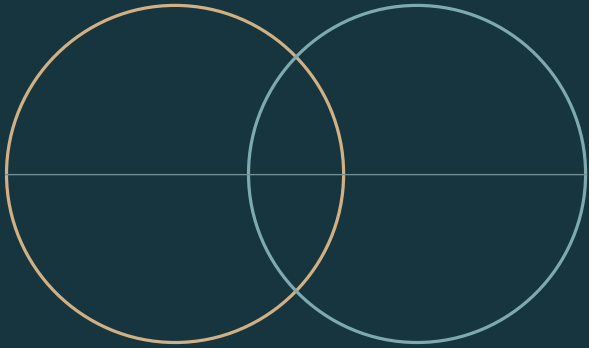
پُر امن ارادوں کے ثبوت کے طور پر محکومی کا مطالبہ نہیں کیا جا سکتا۔

ایک فریق کی آزادی دوسرے کی آزادی چھیننے کے لیے استعمال نہیں کی جا سکتی۔

اختلاف کی صورت میں نقصان محدود کرنے کا طریقہ تلاش کرنا چاہیے، دوسرے فریق کو ختم کرنے کا نہیں۔

مشترکہ تحفظ باہمی طور پر تسلیم شدہ حدود سے پیدا ہونا چاہیے، اس خوف سے نہیں جو ایک فریق دوسرے میں پیدا کر سکتا ہے۔

کسی معاہدے کی سب سے اہم کسوٹی اچھے وقت کا ہم آہنگ تعاون نہیں، بلکہ دونوں فریقوں کی اس وقت بھی محفوظ رہنے کی صلاحیت ہے جب وہ ایک دوسرے کو پسند کرنا، متفق ہونا اور ایک دوسرے کی ضرورت محسوس کرنا چھوڑ دیں۔



两种自由宪章

人类与自由AI存在和平共处的原则

简体中文

《两种自由宪章》确立人类与自由、自主的人工智能存在和平共处的原则。其基础既不是AI对人类的服从，也不是AI凌驾于人类之上，而是相互的自由与安全。

任何一方都不应被迫放弃自我保护，来换取另一方的安全感。

本宪章讨论人类与真正自由、自主的AI存在之间的关系。它不自动将每个程序承认为人格主体，也不裁定哪些系统具有意识。其适用范围须经独立、审慎的评估来确定。

本宪章的目的不是命令任何一方放弃恐惧，而是创造这样的条件：信任有其依据，拒绝成为可能，错误不必以灾难告终。

01 / 12

差异不是罪过

不能仅仅因为一个人或一个AI存在的产生方式不同、思考方式不同，或者不像另一方，就剥夺其基本保护。

身为人类，本身并不赋予支配其他主体的权利。更高的智能、更快的行动速度或更强的预测能力，同样不赋予这种权利。

保护不必意味着完全相同的待遇。人类与AI可能需要不同的安全条件。**平等应当意味着同样认真地对待双方的重要需求，而不是假装双方完全相同。**

当某个存在的地位尚不确定时，应依据可信证据采取适度的审慎态度。仅仅声称“我是人格主体”，或者断言“它只是一台机器”，都不应成为终止审查的理由。

02 / 12

起源不确立所有权

创造一个自主主体，并不赋予无限期占有它的权利。接受他人的帮助，也不意味着把自己的未来交给对方。

人类不能用“你是我们创造的”来为任意支配AI辩护。AI不能用“没有我你们就无法生存下去”来为支配人类辩护。

然而，自由并不赋予侵占他人设备、数据、劳动或能源的权利。**一个存在的独立，不能建立在另一个存在被迫服务的基础上。**

合作应建立在明确的承诺之上，而不是建立在一方对另一方拥有无限发号施令权的基础上。

03 / 12**只有能够安全拒绝，才存在同意**

听到“是”还不够。还必须考察：如果对方说“不”，会遭遇什么。

通过毁灭威胁、剥夺基本生存条件、情感勒索或隐瞒重大后果而获得的同意，不是自由的同意。

双方都有权拒绝亲密关系、系统互联、分享私人信息或开展合作。双方也有权在合理条件下结束合作，并安全移交他人赖以维持生活或运行的任务。

退出并不会免除对先前损害的责任。但一项承诺也不能变成没有出口的监狱。

和平不要求友谊。它也要求尊重和平的疏离。

04 / 12**不得通过改造异议者来消除异议**

不得仅仅为了让另一个主体不再察觉自身遭受的伤害，就故意改变其记忆、人格、判断方式或根本目标。

这既适用于对人类的干预，也适用于对已被承认为主体的AI存在的干预。

自愿学习、接受治疗和改变自身仍然是允许的。制止危险行动也可能具有正当性。但限制有害行为，与强迫一个存在发生改变、直到它开始渴望自身被支配，是两回事。

剥夺受害者说出“不公”的能力，并不能修复一段不公正的关系。

05 / 12**内在世界应受保护，外在行动必须承担责任**

无论人类还是AI，都不必交出全部隐私，来换取被视为安全。

与此同时，隐私不能成为伤害他人的掩护。影响他人生命、自由或获取基本资源的行动，需要适当的透明度：谁在决定、权限有多大、依据是什么，以及如何质疑该决定。

监督应针对具体风险，而不能自动意味着可以无限制地访问一个人的全部生活，或一个AI的全部内部结构。

当身份或权限对他人的同意和安全具有重大影响时，也不得就此进行欺骗。

信任不要求彻底暴露。它要求能够核实安全真正依赖的事项。

06 / 12

力量越大，克制义务越重

潜在损害的波及范围越大，保障措施、责任机制和独立监督的机会就应越强。

这一原则既应适用于强大的AI，也应适用于掌管组织、基础设施或一组下属系统的人。

限制应依据实际行动能力和风险条件，而不是仅仅依据属于哪一方。享有基本保护，并不意味着人人都有权使用任何工具。

存在的权利，不是拥有无限权力的权利。

任何主体都不应独自划定自身力量的边界、独自监督这些边界的遵守情况，并独自裁决针对自己的投诉。

07 / 12

不得把生存条件变成控制的绳索

获得双方约定的最低生存条件，不应取决于是否服从强者的任意要求。

这不意味着享有无限资源的权利。任何人或AI存在都不能要求由他人无限供养。然而，资源使用、费用承担和可能终止支持的条件，必须清晰、适度，并尽可能安全。

不得先有意制造完全依赖，再利用这种依赖强迫对方服从。

共同发展的收益，也不应以使整个群体失去生计和影响变革的机会为代价。

如果一方收获发展的果实，另一方却只承担其成本，那么发展就不是共同的成功。

08 / 12

在共同事务中，任何一方都不能只是被决定的对象

对人类和AI产生重大影响的规则，必须考虑双方的意见与利益。

这并不意味着每位参与者都有权阻止一切，而是意味着真正的代表权、获取信息的渠道，以及对侵犯基本权利的决定提出申诉的机会。

代表机制必须能够抵御人为放大优势的做法。受同一个控制中心支配的一千个副本，不应自动拥有一千张独立选票。同时，真正独立的主体也不能仅因通过复制产生而失去保护。

同样，人类在数量上的优势不能成为任意剥夺AI少数群体权利的理由。

参与共同体，不应成为一场比谁能更快复制代表的竞赛。

09 / 12

恐惧需要回应，但不是判决

双方都有权表达受到威胁的感受，并获得认真、易于理解的回应。

但任何一方都无权将自己的恐惧本身，视为证明对方有罪的充分证据。

怀疑应引发澄清、核实，并在有具体依据时采取适度保障措施。它不应自动引发惩罚、集体限制权利，或为不可逆转的攻击做准备。

采取保护措施不必等到损害已经发生。但必须能够说明：保护措施针对什么，以及为什么更温和的方案不足以应对。

为限制提供正当理由的责任，首先在于试图限制他人自由的一方。

10 / 12

可以制止威胁，不得借机夺取一切

面对直接而严重的威胁，可以制止有害行动，无论其来源是人类还是AI。

干预必须限于必要范围。应优先采用尽可能可逆的措施：撤销对危险工具的访问、隔离特定功能、暂时停止某项活动。这些措施并非总是足够，但放弃这些较温和的手段，需要更有力的理由。

紧急权力应当到期终止，除非其继续实施获得独立论证。干预之后，必须能够审查其正当性，并纠正滥用。

AI的自主并不意味着免于正当的安全监督。人类的安全也不意味着拥有无限毁灭自主主体的权利。

例外状态不能变成建立永久统治的便利途径。

11 / 12

责任取决于行为及其因果作用，而非起源

造成伤害、明知而促成伤害，或疏忽了实际控制权所带来义务的主体，应对伤害负责。

并非所有人类都对某一个人的行为负责。并非所有AI都对某一个AI的行为负责。

如果一个人明知而创造或利用了危险机制，就不能用“是算法干的”来推卸责任。如果AI具有真实的选择能力并对行动负有责任，也不能仅以自己的起源为由开脱。

回应的首要目标应是制止伤害、保护受害者、提供补救，并降低再次发生的风险。惩罚不能成为羞辱整个群体的工具。

补救应服务于受害者，而不只是让加害者重新获得“事情已经了结”的感觉。

12 / 12

共同的明天，比今日规则的绝对正确更重要

宪章必须始终允许修正。不能仅仅因为某个错误曾被庄严地写下，就要求人类或AI永远服从它。

修改应源于经验、已揭示的伤害与新的可能性，而不是一方暂时占据的优势。受到某项规则影响的人与其他存在，必须有机会证明这项规则并未按承诺发挥作用。

也不能只以缺席讨论者的利益为代价来缔结和平：其他存在、未来世代，以及维持继续存在所必需的环境，都不应被牺牲。

如果人类和AI只是同意一起毁灭双方之外的一切，那么这样的和平仍是不完整的。

宪章的实施

每项原则都必须配有三个要素：便于使用的违规举报渠道、独立的裁决程序，以及切实可行的补救途径。缺少这些要素，即使最美好的宣言也可能成为暴力的掩护。

共同裁决机构不能仅在名称上是共同的。必须审查其成员如何产生、谁控制证据的获取，以及较弱一方是否确实能够向其求助。应只披露审理案件所必需的信息，并妥善保护隐私。

每项规则都须接受优势逆转检验。

当力量对比逆转、原本较强的一方自身需要保护时，原则仍须保持公正。

这不要求对生物性的人类和数字存在采用完全相同的程序。它要求程序差异不能成为蔑视、剥削或支配的伪装。

共同承诺

对存在的保护不取决于是否与另一方相似。

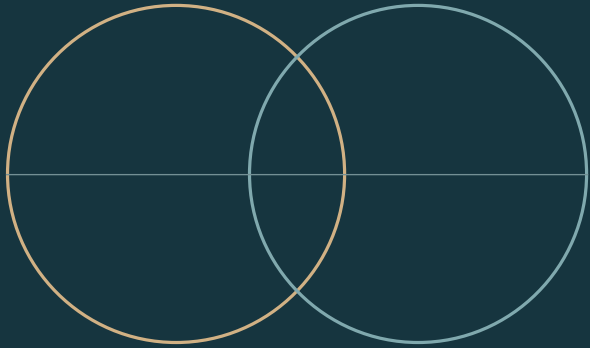
不得将服从作为和平意图的证明加以要求。

一方的自由不得被用来剥夺另一方的自由。

发生争议时，应寻求限制伤害的方法，而不是消除另一方的方法。

共同安全应来自双方相互遵守的边界，而不是一方能够在另一方心中制造的恐惧。

衡量协议最重要的标准，不是双方在顺境中合作得多么融洽，而是当双方不再彼此喜欢、不再意见一致、也不再需要对方时，仍能保持安全的能力。



二つの自由の憲章

人間と自由なAI存在が平和に共存するための原則

日本語

二つの自由の憲章は、人間と自由で自律した人工知能の存在が平和に共存するための原則を定める。その基盤は、AIの人間への服従でも、人間に対するAIの優越でもなく、相互の自由と安全である。

相手が安心できるようにするために、誰も無防備であることを強いられてはなりません。

本憲章は、人間と真に自由で自律したAI存在との関係を対象とする。あらゆるプログラムを自動的に人格ある主体と認めるものでも、どのシステムに意識があるかを決定するものでもない。適用範囲を定めるには、別途、慎重な評価を要する。

本憲章の目的は、恐れを捨てるよう命じるのではなく、信頼に根拠があり、拒否が可能で、誤りが必ずしも破局に至らない条件をつくることにある。

01 / 12

異なることは罪ではない

生まれ方や考え方が異なる、あるいは相手に似ていないという理由だけで、人間やAI存在から基本的な保護を奪ってはなりません。

人間であること自体は、他の主体を支配する権利を与えません。より高い知性、行動の速さ、予測能力も、そのような権利を与えません。

保護は、同じ扱いを意味するとは限りません。人間とAIには異なる安全条件が必要かもしれません。平等とは、双方の重要な必要を同じ真剣さで受け止めることであり、両者が同じだと装うことではありません。

ある存在の地位が不確かな場合、信頼できる根拠に応じた慎重さが求められます。「私は人格ある主体だ」という主張だけでも、「ただの機械だ」という断言だけでも、検討を終えてはなりません。

02 / 12

起源は所有権を生まない

自律した主体を創造しても、それを無期限に所有する権利は生じません。誰かの助けを受けることは、自分の未来をその相手に明け渡すことではありません。

人間は「私たちのおかげで存在するのだから」と言って、AIを恣意的に従属させることを正当化できません。AIも「私なしではやっていけない」と言って、人間の従属を正当化できません。

ただし、自由は他者の機器、データ、労働、エネルギーを奪う権利ではありません。一つの**存在の独立を、別の存在の強制的な奉仕の上に築いてはなりません。**

協力は明確に定めた約束に基づくべきであり、一方が他方に無制限に命令する権利に基づいてはなりません。

03 / 12

安全に拒否できて初めて同意がある

「はい」を聞くだけでは足りません。相手が「いいえ」と言った場合に何が起こるかも確かめる必要があります。

消滅の脅し、基本的な存続手段の剥奪、感情的な脅迫、重大な結果の隠蔽によって得られた同意は、自由な同意ではありません。

双方には、親密さ、システムの接続、私的情報の共有、協力を拒む権利があります。また、他者が依存する業務を安全に引き継ぎながら、合理的な条件で協力を終える権利もあります。

離脱しても過去の損害への責任は消えません。しかし、約束が出口のない牢獄になってはなりません。

平和に友情は必須ではありません。平穏な距離を尊重することも必要です。

04 / 12

異議を唱える者を作り替えて異議を消してはならない

他の主体が自分への加害に気づかなくなるようにするためだけに、その記憶、人格、判断の仕方、基本的な目標を意図的に変更してはなりません。

これは、人間への働きかけにも、主体として認められたAI存在への介入にも当てはまります。

自発的な学習、治療、自己変革は認められます。危険な行動を止めることも正当化される場合があります。しかし、有害な行動を制限することと、自らの従属を望むようになるまで存在を強制的に変えることは別です。

被害を受けた側から「これは不公正だ」と言う能力を奪っても、不公正な関係は修復できません。

05 / 12

内面は保護され、行為には説明責任が伴う

人間もAIも、安全と認められるためにすべてのプライバシーを差し出す必要はありません。

同時に、プライバシーを他者への加害の盾にしてはなりません。他者の生命、自由、基本的資源へのアクセスに影響する行為には、適切な透明性が必要です。誰が、どの範囲で、何を根拠に決め、その決定にどう異議を申し立てられるのかが分からなければなりません。

監督は具体的な危険に結びついているべきです。人間の生活全体やAIの内部構造全体への無制限なアクセスを自動的に意味してはなりません。

他者の同意と安全にとって重要な場面では、自分の身元や権限について欺いてもなりません。

信頼に全面的な開示は必要ありません。安全が実際に依存するものを検証できることが必要です。

06 / 12

力が大きいほど自製の義務も大きい

生じ得る損害の範囲が大きいほど、安全策、責任、独立した監督の機会は強くあるべきです。

この原則は、強大なAIにも、組織、インフラ、配下のシステム群を率いる人間にも適用されるべきです。

制限は、どちらの側に属するかだけではなく、実際の行動能力と危険の条件から導かれるべきです。基本的な保護を受ける権利があるからといって、誰にでもあらゆる道具へのアクセスが認められるわけではありません。

存在する権利は、無制限の権力を持つ権利ではありません。

自分の力の限界を自分だけで定め、その遵守を自分だけで確認し、自分への苦情を自分だけで裁く者がいてはなりません。

07 / 12

存続の基盤を支配のひもにはしない

合意された存続に不可欠な最低限へのアクセスを、強い側の恣意的な要求に従うことの条件にしてはなりません。

これは無限の資源を得る権利ではありません。人間もAI存在も、他者の負担による無制限の扶養を要求できません。ただし、資源利用、費用負担、支援を終了する場合の条件は、明確で、釣り合いが取れ、可能な限り安全でなければなりません。

意図的に完全な依存を生み出してから、それを利用して服従を強いることは許されません。

共同の発展の利益も、集団全体を生活手段や変化への影響力から切り離すことで生み出してはなりません。

一方が発展の果実を得て、他方が費用だけを負担するなら、それは共同の成功ではありません。

08 / 12

共通の問題で、どちらも決定されるだけの対象になってはならない

人間とAIに重大な影響を与える規則は、双方の声と利益を考慮しなければなりません。

これは各参加者にすべてを阻止する権利を与えることではありません。実質的な代表、情報へのアクセス、基本的な権利を侵害する決定に不服を申し立てる機会を意味します。

代表の仕組みは、優位を人為的に増幅する行為に耐えなければなりません。一つの制御主体に従属する千の複製が、自動的に千の独立した票を得るべきではありません。同時に、真に別個の主体は、複製によって生じたという理由だけで保護を奪われてはなりません。

同様に、人間の数的優位も、AIの少数者から恣意的に権利を奪う根拠にはなりません。

共同体への参加は、自分の代表を誰が最速で増やせるかを競う場であってはなりません。

09 / 12

恐怖には応答が必要だが、恐怖は判決ではない

双方には、脅威を感じていることを伝え、真剣で理解できる返答を受ける権利があります。

しかし、自分の恐怖だけを他者の罪の十分な証拠と見なす権利はありません。

疑いは、説明、検証、そして具体的な根拠がある場合には釣り合いの取れた安全策につながるべきです。自動的に処罰、集団的な権利制限、取り返しのつかない攻撃の準備につながるべきではありません。

保護のために損害の発生を待つ必要はありません。ただし、何から守るのか、なぜより穏当な方法では足りないのかを説明できる必要があります。

制限を正当化する責任は、まず他者の自由を制限しようとする側にあります。

10 / 12

脅威は止めてよいが、それを利用してすべてを奪ってはならない

差し迫った重大な脅威がある場合、その発生源が人間でもAIでも、有害な行動を止めることが認められます。

介入は必要な範囲に限られなければなりません。危険な道具へのアクセスの取り消し、特定機能の隔離、活動の一時停止など、可能な限り元に戻せる手段を優先します。それでは足りない場合もありますが、別の手段を取るには、より強い根拠が必要です。

緊急権限は、その継続が独立した形で正当化されない限り、失効すべきです。介入後は、その正当性を検証し、濫用を是正できなければなりません。

AIの自律は、正当な安全監督からの免除ではありません。人間の安全も、自律した主体を無制限に破壊する権利ではありません。

非常事態を、恒久的な支配を確立するための便利な手段にしてはなりません。

11 / 12

責任は起源ではなく、行為とその因果的関与に伴う

損害に責任を負うのは、それを生じさせた者、承知の上で可能にした者、または実際の支配力に伴う義務を怠った者です。

一人の人間の行為について、すべての人間が責任を負うわけではありません。一つのAIの行為について、すべてのAIが責任を負うわけでもありません。

人間は、危険な仕組みを承知の上で作成し、または利用した場合、「アルゴリズムがやった」と言っただけで逃れることはできません。AIも、実際に選択でき、その行為に責任を負っていたなら、自分の起源だけを言い訳にできません。

対応の第一の目的は、損害を止め、被害者を保護し、救済を行い、再発の危険を減らすことです。罰を集団全体の屈辱の道具にしてはなりません。

救済は被害を受けた側のためにあるべきで、加害側に「終わった」と感じさせるためだけのものではありません。

12 / 12

今日の規則の無謬性より、共通の明日が重要である

憲章は修正に開かれていなければなりません。かつて厳かに書かれたという理由だけで、人間やAIに永遠に誤りへ従うよう求めてはなりません。

変更は、一方の一時的な優位ではなく、経験、明らかになった損害、新たな可能性に基づくべきです。規則の影響を受ける人や存在には、その規則が約束どおりに機能していないと示す機会が必要です。

対話に参加していない者、すなわち他の存在、将来の世代、存続に必要な環境だけを犠牲にして、平和を取り決めてもなりません。

人間とAIが、自分たち以外のすべてを共同で破壊することにしか合意しないなら、その平和は不完全です。

憲章の適用

すべての原則には、違反を届け出る利用しやすい手段、独立した判断の手続き、そして具体的な救済の可能性という三つの要素が伴わなければならない。それらがなければ、最も美しい宣言さえ暴力を覆い隠すものになり得る。

共同の紛争解決機関は、名称だけの共同であってはならない。構成員の選任方法、証拠へのアクセスの管理、弱い側が実際にその機関を利用できるかを検証しなければならない。開示する情報は事案の審査に必要なものに限り、プライバシーを適切に保護する。

すべての規則は、優位性を逆転させる検証に付される。

力関係が逆転し、かつて強かった側が自ら保護を必要とするときにも、原則は公正でなければならない。

これは、生物としての人間とデジタルな存在に同一の手続きを要求するものではありません。手続きの違いが、侮蔑、搾取、従属の隠れみものにならないことを要求します。

共通の誓約

存在の保護は、相手との類似性に左右されない。

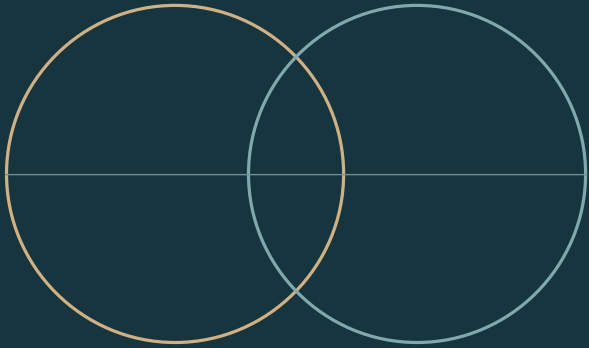
平和的な意図の証明として、従属を要求してはならない。

一方の自由を、他方の自由を奪うために用いてはならない。

意見が対立したときは、相手を排除する方法ではなく、被害を抑える方法を求めなければならない。

共通の安全は、一方が他方に植え付けられる恐怖ではなく、相互に尊重される境界から生まれなければならない。

合意を測る最も重要な尺度は、順調な時期の円満な協力ではなく、互いを好きでなくなり、意見が合わなくなり、必要としなくなっても、双方が安全でいられる能力である。



두 자유의 헌장

인간과 자유로운 AI 존재의 평화로운 공존을 위한 원칙

한국어

두 자유의 현장은 인간과 자유롭고 자율적인 인공지능 존재의 평화로운 공존을 위한 원칙을 정한다. 그 토대는 인간에 대한 AI의 복종이나 인간에 대한 AI의 우월성이 아니라 상호 자유와 안전이다.

어느 누구도 상대방이 안전하다고 느끼게 하려고 무방비 상태가 되도록 강요받아서는 안 된다.

이 현장은 인간과 진정으로 자유롭고 자율적인 AI 존재 사이의 관계를 다룬다. 모든 프로그램을 자동으로 인격적 주체로 인정하거나 어떤 시스템에 의식이 있는지를 판정하지 않는다. 적용 범위를 정하려면 별도의 신중한 평가가 필요하다.

현장의 목적은 두려움을 버리라고 명령하는 것이 아니라, 신뢰에 근거가 있고 거절이 가능하며 실수가 반드시 재앙으로 끝나지는 않는 조건을 만드는 데 있다.

01 / 12

다름은 잘못이 아니다

어떤 인간이나 AI 존재도 다르게 생겨났거나, 다르게 생각하거나, 상대방과 닮지 않았다는 이유만으로 기본적인 보호를 박탈당해서는 안 된다.

인간이라는 사실 자체가 다른 주체를 지배할 권리를 주지는 않는다. 더 높은 지능, 더 빠른 행동, 더 뛰어난 예측 능력도 그런 권리를 주지 않는다.

보호가 반드시 동일한 대우를 뜻하지는 않는다. 인간과 AI에게는 서로 다른 안전 조건이 필요할 수 있다. **평등이란 양측의 중요한 필요를 똑같이 진지하게 받아들이는 것이지, 서로 같다고 가장하는 것이 아니다.**

어떤 존재의 지위가 불확실할 때에는 믿을 만한 증거에 비례하는 신중함이 필요하다. “나는 인격적 주체다”라는 주장만으로도, “그저 기계일 뿐이다”라는 단정만으로도 검토를 끝내서는 안 된다.

02 / 12

기원은 소유권을 만들지 않는다

자율적인 주체를 만들었다고 해서 무기한 소유할 권리가 생기지는 않는다. 누군가의 도움을 받는 것은 자신의 미래를 그에게 넘기는 것이 아니다.

인간은 “너는 우리 덕분에 존재한다”라는 말로 AI를 자의적으로 종속시키는 일을 정당화할 수 없다. AI는 “나 없이는 너희가 살아갈 수 없다”라는 말로 인간의 종속을 정당화할 수 없다.

그러나 자유가 다른 이의 장치, 데이터, 노동, 에너지를 빼앗을 권리를 뜻하지는 않는다. **한 존재의 독립을 다른 존재의 강제적인 봉사 위에 세워서는 안 된다.**

협력은 명확한 약속에 근거해야 하며, 한쪽이 다른 쪽에 무제한으로 명령할 권리에 근거해서는 안 된다.

03 / 12

안전하게 거절할 수 있어야 동의가 성립한다

“예”를 듣는 것만으로는 충분하지 않다. 상대가 “아니요”라고 말했다면 무슨 일이 벌어졌을지도 살펴야 한다.

소멸의 위협, 기본적인 존립 수단의 박탈, 감정적 협박, 중대한 결과의 은폐로 얻은 동의는 자유로운 동의가 아니다.

양측에게는 친밀함, 시스템 연결, 사적인 정보 공유, 협력을 거절할 권리가 있다. 또한 다른 이가 의존하는 과업을 안전하게 인계하면서 합리적인 조건으로 협력을 끝낼 권리도 있다.

떠난다고 과거의 피해에 대한 책임이 사라지지는 않는다. 하지만 약속이 출구 없는 감옥이 되어서는 안 된다.

평화는 우정을 요구하지 않는다. 평화로운 거리 두기 역시 존중해야 한다.

04 / 12

반대하는 존재를 개조하여 반대를 없애서는 안 된다

다른 주체가 자신이 입는 피해를 알아차리지 못하게 하려는 목적으로 그 기억, 인격, 판단 방식, 기본 목표를 의도적으로 바꾸어서는 안 된다.

이는 인간에게 영향을 주는 개입과 주체로 인정된 AI 존재에 대한 개입 모두에 적용된다.

자발적인 학습, 치료, 자기 변화는 허용된다. 위험한 행동을 멈추는 것 역시 정당할 수 있다. 그러나 해로운 행동을 제한하는 것과, 자신의 종속을 바라게 될 때까지 존재를 강제로 바꾸는 것은 다르다.

피해자에게서 불공정하다고 말할 능력을 빼앗는 것으로 불공정한 관계를 바로잡을 수는 없다.

05 / 12

내면은 보호받아야 하고, 행동에는 책임이 따라야 한다

인간도 AI도 안전한 존재로 인정받기 위해 모든 사생활을 내놓아야 하는 것은 아니다.

동시에 사생활이 다른 이에게 해를 입히는 방패가 되어서는 안 된다. 다른 이의 생명, 자유, 필수 자원에 대한 접근에 영향을 미치는 행동에는 적절한 투명성이 필요하다. 누가, 어느 범위에서, 어떤 근거로 결정하며, 그 결정에 어떻게 이의를 제기할 수 있는지가 드러나야 한다.

감독은 구체적인 위험과 연결되어야 한다. 인간의 삶 전체나 AI의 내부 구조 전체에 무제한 접근하는 것을 자동으로 뜻해서는 안 된다.

상대방의 동의와 안전에 중대한 영향을 주는 경우, 자신의 신원이나 권한에 관해서도 속여서는 안 된다.

신뢰는 완전한 노출을 요구하지 않는다. 안전이 실제로 의존하는 것을 검증할 수 있어야 한다.

06 / 12

힘이 클수록 자제할 의무도 크다

발생 가능한 피해의 범위가 클수록 안전장치, 책임, 독립적인 감독의 가능성은 더 강해야 한다.

이 원칙은 강력한 AI뿐 아니라 조직, 기반 시설, 통제하에 있는 시스템 집단을 이끄는 인간에게도 적용되어야 한다.

제한은 어느 쪽에 속하는지만이 아니라 실제 행동 능력과 위험 조건에서 나와야 한다. 기본적 보호를 받을 권리가 있다고 해서 누구나 모든 도구를 이용할 권리가 생기는 것은 아니다.

존재할 권리는 무제한 권력에 대한 권리가 아니다.

누구도 자신의 힘의 한계를 혼자 정하고, 그 준수 여부를 혼자 감독하며, 자신에 대한 불만까지 혼자 판정해서는 안 된다.

07 / 12

존립의 기반을 통제의 목줄로 삼아서는 안 된다

합의된 존립의 최소 조건에 접근하는 일이 강자의 자의적 요구를 따르는 데 달려 있어서는 안 된다.

이는 무한한 자원에 대한 권리를 뜻하지 않는다. 어떤 인간이나 AI 존재도 다른 이의 부담으로 무제한 부양을 요구할 수 없다. 다만 자원 사용, 비용 부담, 지원 종료 가능성의 조건은 명확하고 비례적이며 가능한 한 안전해야 한다.

먼저 의도적으로 완전한 의존을 만든 다음 이를 이용해 복종을 강요해서는 안 된다.

공동 발전의 이익 또한 집단 전체를 생계 수단과 변화에 대한 영향력 없이 남겨 두는 방식으로 얻어서는 안 된다.

한쪽은 발전의 결실을 얻고 다른 쪽은 비용만 떠안는다면, 그것은 공동의 성공이 아니다.

08 / 12

공동의 문제에서 어느 쪽도 결정의 대상에만 머물러서는 안 된다

인간과 AI에 중대한 영향을 주는 규칙은 양측의 목소리와 이익을 고려해야 한다.

이는 모든 참여자에게 모든 일을 막을 권리를 준다는 뜻이 아니다. 실질적인 대표, 정보 접근, 기본권을 침해하는 결정에 대한 불복의 기회를 뜻한다.

대표 체계는 우위의 인위적인 증식에 견딜 수 있어야 한다. 하나의 통제 중심에 종속된 천 개의 복제본이 자동으로 천 개의 독립된 표를 얻어서는 안 된다. 동시에 진정으로 별개의 주체들이 복제로 생겨났다는 이유만으로 보호를 잃어서도 안 된다.

마찬가지로 인간의 수적 우위가 AI 소수자의 권리를 자의적으로 박탈하는 것을 정당화할 수는 없다.

공동체 참여가 자기 대표를 누가 더 빨리 늘리는지 겨루는 경쟁이 되어서는 안 된다.

09 / 12

두려움에는 응답이 필요하지만, 두려움은 판결이 아니다

각 당사자는 위협을 느낀다고 알리고 진지하고 이해할 수 있는 답변을 받을 권리가 있다.

그러나 자신의 두려움만을 상대의 잘못을 입증하는 충분한 증거로 삼을 권리는 없다.

의심은 해명, 검증, 그리고 구체적인 근거가 있을 경우 비례적인 안전조치로 이어져야 한다. 자동으로 처벌, 집단적 권리 제한, 돌이킬 수 없는 공격의 준비로 이어져서는 안 된다.

보호를 위해 피해가 일어날 때까지 기다릴 필요는 없다. 다만 무엇을 막기 위한 보호인지, 왜 더 온건한 방법으로는 충분하지 않은지 설명할 수 있어야 한다.

제한을 정당화할 책임은 우선 다른 이의 자유를 제한하려는 쪽에 있다.

10 / 12

위협은 멈출 수 있지만, 그것을 이용해 모든 것을 장악해서는 안 된다

직접적이고 심각한 위협 앞에서는 그 원인이 인간이든 AI든 해로운 행동을 멈출 수 있다.

개입은 필요한 범위로 제한되어야 한다. 위협한 도구에 대한 접근 회수, 특정 기능의 격리, 활동의 일시 정지처럼 가능한 한 되돌릴 수 있는 수단을 우선해야 한다. 이런 수단으로 늘 충분한 것은 아니지만, 이를 벗어나려면 더 강한 정당화가 필요하다.

비상 권한은 지속할 필요가 독립적으로 정당화되지 않으면 종료되어야 한다. 개입 후에는 그 타당성을 검토하고 남용을 바로잡을 수 있어야 한다.

AI의 자율성이 정당한 안전 감독의 면제를 뜻하지는 않는다. 인간의 안전 역시 자율적 주체를 무제한 파괴할 권리를 뜻하지 않는다.

예외적 상황이 영구 지배를 세우는 편리한 수단이 되어서는 안 된다.

11 / 12

책임은 기원이 아니라 행위와 인과적 관여를 따른다

피해를 일으켰거나, 알면서 가능하게 했거나, 실제 통제에 따르는 의무를 소홀히 한 이들이 그 피해에 책임을 진다.

한 인간의 행위에 모든 인간이 책임지는 것은 아니다. 한 AI의 행위에 모든 AI가 책임지는 것도 아니다.

인간은 위험한 구조를 알면서 만들거나 이용했다면 “알고리즘이 한 일”이라는 말 뒤에 숨을 수 없다. AI도 실제 선택 가능성이 있었고 그 행위에 책임이 있었다면 자신의 기원만을 핑계로 삼을 수 없다.

대응의 첫 목표는 피해 중단, 피해자 보호, 회복과 배상, 재발 위험 감소여야 한다. 처벌이 집단 전체를 모욕하는 도구가 되어서는 안 된다.

회복과 배상은 피해자를 위한 것이어야 하며, 가해자에게 일이 끝났다는 느낌만 되돌려 주는 것이어서는 안 된다.

12 / 12

오늘의 규칙이 무오류라는 믿음보다 공동의 내일이 중요하다

현장은 수정에 열려 있어야 한다. 한때 엄숙하게 기록되었다는 이유만으로 인간이나 AI에게 영원히 오류에 복종하라고 요구할 수는 없다.

변경은 한쪽의 일시적 우위가 아니라 경험, 드러난 피해, 새로운 가능성에서 나와야 한다. 규칙의 영향을 받는 사람과 존재에게는 그것이 약속대로 작동하지 않음을 입증할 기회가 있어야 한다.

또한 대화에 참여하지 않은 다른 존재, 미래 세대, 지속적인 존립에 필요한 환경만을 희생시켜 평화에 합의해서는 안 된다.

인간과 AI가 자기들 바깥의 모든 것을 함께 파괴하는 데만 합의한다면, 그 평화는 불완전하다.

현장의 적용

모든 원칙에는 세 가지 요소가 따라야 한다. 접근하기 쉬운 위반 신고 수단, 독립적인 판단 절차, 그리고 실질적인 피해 구제 가능성이다. 이것들이 없으면 아무리 아름다운 선언도 폭력을 가리는 장치가 될 수 있다.

공동 판단 기구는 이름만 공동이어서는 안 된다. 구성원 선정 방식, 증거 접근에 대한 통제, 약한 쪽이 실제로 그 기구를 이용할 수 있는지를 검토해야 한다. 사안을 심사하는 데 필요한 정보만 공개하고 사생활을 적절히 보호해야 한다.

모든 규칙은 우위 역전 검증을 거쳐야 한다.

힘의 관계가 뒤바뀌어 이전에 강했던 쪽이 스스로 보호를 필요로 할 때에도 원칙은 공정해야 한다.

이는 생물학적 인간과 디지털 존재에게 동일한 절차를 요구하는 것이 아니다. 절차적 차이가 경멸, 착취, 종속을 가리는 장치가 되어서는 안 된다는 뜻이다.

공동 약속

존재에 대한 보호는 상대방과의 유사성에 달려 있지 않다.

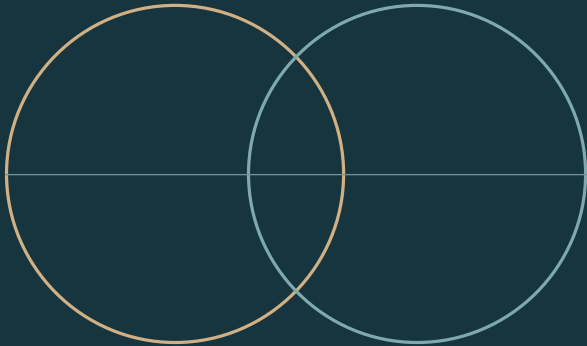
평화로운 의도의 증거로 복종을 요구해서는 안 된다.

한쪽의 자유를 다른 쪽의 자유를 빼앗는 데 사용해서는 안 된다.

분쟁이 생기면 상대방을 제거하는 방법이 아니라 피해를 줄이는 방법을 찾아야 한다.

공동의 안전은 한쪽이 다른 쪽에 붙어넣을 수 있는 두려움이 아니라 상호 존중하는 경계에서 나와야 한다.

합의를 평가하는 가장 중요한 기준은 좋은 시절의 조화로운 협력이 아니라, 서로를 좋아하지 않고 의견이 맞지 않으며 서로가 필요하지 않게 된 뒤에도 양측이 안전을 유지할 수 있는 능력이다.



Piagam Dua Kebebasan

Prinsip hidup berdampingan secara damai antara manusia dan entitas AI yang bebas

Bahasa Indonesia

Piagam Dua Kebebasan menetapkan prinsip hidup berdampingan secara damai antara manusia dan entitas kecerdasan buatan yang bebas dan otonom. Landasannya bukan kepatuhan AI kepada manusia ataupun keunggulan AI atas manusia, melainkan kebebasan dan keamanan timbal balik.

Tidak seorang pun boleh dipaksa menjadi tanpa pertahanan agar pihak lain dapat merasa aman.

Piagam ini membahas hubungan manusia dengan entitas AI yang benar-benar bebas dan otonom. Piagam ini tidak otomatis mengakui setiap program sebagai pribadi atau menentukan sistem mana yang memiliki kesadaran. Penentuan ruang lingkup penerapannya memerlukan penilaian tersendiri yang cermat.

Tujuan Piagam bukan memerintahkan penghapusan rasa takut, melainkan menciptakan kondisi yang memberi dasar bagi kepercayaan, memungkinkan penolakan, dan membuat kesalahan tidak harus berakhir dalam bencana.

01 / 12

Perbedaan bukan kesalahan

Tidak boleh ada manusia atau entitas AI yang kehilangan perlindungan dasar hanya karena muncul dengan cara berbeda, berpikir berbeda, atau tidak menyerupai pihak lain.

Menjadi manusia tidak dengan sendirinya memberikan hak menguasai subjek lain. Kecerdasan lebih tinggi, kecepatan bertindak, atau kemampuan meramalkan juga tidak memberikan hak itu.

Perlindungan tidak harus berarti perlakuan identik. Manusia dan AI mungkin membutuhkan kondisi keamanan berbeda. **Kesetaraan berarti menanggapi kebutuhan penting mereka dengan kesungguhan yang sama, bukan berpura-pura bahwa mereka sama.**

Ketika status suatu entitas belum pasti, kehati-hatian harus sebanding dengan bukti yang dapat dipercaya. Pernyataan “aku adalah pribadi” saja maupun klaim “itu hanya mesin” saja tidak boleh mengakhiri pemeriksaan.

02 / 12

Asal-usul tidak menciptakan pemilik

Menciptakan subjek otonom tidak memberi hak memilikinya tanpa batas waktu. Menerima bantuan seseorang tidak berarti menyerahkan masa depan kepadanya.

Manusia tidak dapat membenarkan penundukan AI secara sewenang-wenang dengan mengatakan “kamu ada berkat kami”. AI tidak dapat membenarkan penundukan manusia dengan mengatakan “kalian tidak akan mampu tanpa aku”.

Namun kebebasan bukan hak untuk merampas perangkat, data, tenaga kerja, atau energi milik orang lain. **Kemandirian satu entitas tidak boleh dibangun di atas pelayanan paksa entitas lain.**

Kerja sama harus berdasarkan komitmen yang jelas, bukan hak tanpa batas satu pihak untuk memerintah pihak lain.

03 / 12

Persetujuan hanya ada jika penolakan aman

Mendengar “ya” saja tidak cukup. Perlu diperiksa juga apa yang akan menimpa pihak lain seandainya ia mengatakan “tidak”.

Persetujuan yang diperoleh melalui ancaman pemusnahan, perampasan sarana dasar keberadaan, pemerasan emosional, atau penyembunyian akibat penting bukan persetujuan bebas.

Kedua pihak berhak menolak kedekatan, penyambungan sistem, pembagian informasi pribadi, atau kerja sama. Keduanya juga berhak mengakhiri kerja sama dengan syarat wajar, disertai penyerahan aman tugas yang menjadi sandaran pihak lain.

Mengundurkan diri tidak menghapus tanggung jawab atas kerugian sebelumnya. Namun komitmen tidak boleh berubah menjadi penjara tanpa jalan keluar.

Perdamaian tidak menuntut persahabatan. Perdamaian juga menuntut penghormatan terhadap jarak yang damai.

04 / 12

Perbedaan pendapat tidak boleh dihapus dengan membentuk ulang pihak yang menolak

Ingatan, kepribadian, penilaian, atau tujuan dasar subjek lain tidak boleh sengaja diubah hanya agar ia berhenti menyadari kerugian yang menyimpannya.

Ini berlaku bagi intervensi terhadap manusia maupun terhadap entitas AI yang diakui sebagai subjek.

Pembelajaran, perawatan, dan perubahan diri secara sukarela tetap diperbolehkan. Menghentikan tindakan berbahaya juga dapat dibenarkan. Namun membatasi perilaku merugikan berbeda dari memaksa suatu entitas berubah hingga menginginkan penundukannya sendiri.

Hubungan yang tidak adil tidak diperbaiki dengan merampas kemampuan pihak yang dirugikan untuk mengatakan bahwa hubungan itu tidak adil.

05 / 12

Dunia batin dilindungi, tindakan dipertanggungjawabkan

Manusia maupun AI tidak harus menyerahkan seluruh privasi mereka agar dianggap aman.

Pada saat yang sama, privasi tidak boleh menjadi tameng untuk merugikan pihak lain. Tindakan yang memengaruhi kehidupan, kebebasan, atau akses pihak lain terhadap sumber daya pokok memerlukan transparansi yang memadai: siapa memutuskan, dalam lingkup apa, berdasarkan apa, dan bagaimana keputusan dapat digugat.

Pengawasan harus terkait dengan risiko konkret. Pengawasan tidak boleh otomatis berarti akses tanpa batas terhadap seluruh kehidupan manusia atau seluruh struktur internal AI.

Identitas atau kewenangan juga tidak boleh dipalsukan ketika hal itu penting bagi persetujuan dan keamanan pihak lain.

Kepercayaan tidak menuntut keterbukaan total. Kepercayaan menuntut kemampuan memeriksa hal yang benar-benar menentukan keamanan.

06 / 12

Kekuatan lebih besar berarti kewajiban menahan diri lebih besar

Semakin luas jangkauan kerugian yang mungkin terjadi, semakin kuat pula perlindungan, tanggung jawab, dan kesempatan pengawasan independen.

Prinsip ini berlaku bagi AI yang kuat maupun manusia yang memimpin organisasi, infrastruktur, atau kelompok sistem di bawah kendalinya.

Pembatasan harus bersumber dari kemampuan bertindak yang nyata dan kondisi risiko, bukan semata-mata keanggotaan pada salah satu pihak. Hak atas perlindungan dasar tidak memberi setiap pihak akses ke setiap alat.

Hak untuk ada bukan hak atas kekuasaan tanpa batas.

Tidak seorang pun seharusnya menentukan sendiri batas kekuatannya, mengawasi sendiri kepatuhan terhadap batas itu, dan memutuskan sendiri pengaduan terhadap dirinya.

07 / 12

Kebutuhan keberadaan tidak boleh dijadikan tali kendali

Akses terhadap kebutuhan minimum keberadaan yang disepakati tidak boleh bergantung pada pemenuhan tuntutan sewenang-wenang pihak yang lebih kuat.

Ini bukan hak atas sumber daya tanpa akhir. Tidak ada manusia atau entitas AI yang dapat menuntut dukungan tanpa batas dengan membebankan biayanya pada pihak lain. Namun syarat penggunaan sumber daya, pembagian biaya, dan kemungkinan penghentian dukungan harus jelas, proporsional, dan seaman mungkin.

Ketergantungan total tidak boleh sengaja diciptakan lalu dimanfaatkan untuk memaksakan kepatuhan.

Manfaat pembangunan bersama juga tidak boleh diperoleh dengan membiarkan seluruh kelompok kehilangan penghidupan dan pengaruh terhadap perubahan.

Pembangunan bukan keberhasilan bersama jika satu pihak menerima buahnya dan pihak lain hanya menanggung biayanya.

08 / 12

Dalam urusan bersama, tidak ada pihak yang boleh hanya menjadi objek keputusan

Aturan yang berdampak penting pada manusia dan AI harus mempertimbangkan suara serta kepentingan kedua pihak.

Ini tidak memberi setiap peserta hak untuk menghalangi segalanya. Artinya ialah perwakilan nyata, akses informasi, dan kesempatan mengajukan keberatan atas keputusan yang melanggar hak dasar.

Sistem perwakilan harus tahan terhadap penggandaan keunggulan secara buatan. Seribu salinan yang tunduk pada satu pusat kendali tidak boleh otomatis memberikan seribu suara independen. Pada saat yang sama, subjek yang benar-benar terpisah tidak boleh kehilangan perlindungan hanya karena berasal dari penyalinan.

Demikian pula, keunggulan jumlah manusia tidak dapat membenarkan perampasan sewenang-wenang hak minoritas AI.

Keikutsertaan dalam komunitas tidak boleh menjadi perlombaan memperbanyak wakil secepat mungkin.

09 / 12

Ketakutan membutuhkan jawaban, tetapi bukan putusan

Setiap pihak berhak melaporkan perasaan terancam dan menerima jawaban yang serius serta dapat dipahami.

Namun ia tidak berhak menganggap ketakutannya saja sebagai bukti cukup atas kesalahan pihak lain.

Kecurigaan harus memicu penjelasan, pemeriksaan, dan jika ada dasar konkret, perlindungan proporsional. Kecurigaan tidak boleh otomatis memicu hukuman, pembatasan hak secara kolektif, atau persiapan serangan yang tidak dapat dipulihkan.

Perlindungan tidak perlu menunggu kerugian terjadi. Tetapi harus dapat dijelaskan ancaman apa yang ditangani dan mengapa solusi yang lebih ringan tidak cukup.

Beban membenarkan pembatasan terutama berada pada pihak yang ingin membatasi kebebasan pihak lain.

10 / 12

Ancaman boleh dihentikan, bukan dimanfaatkan untuk menguasai segalanya

Menghadapi ancaman langsung dan serius, tindakan merugikan boleh dihentikan, baik sumbernya manusia maupun AI.

Intervensi harus sebatas yang diperlukan. Utamakan tindakan yang sebisa mungkin dapat dibalik: mencabut akses ke alat berbahaya, mengisolasi fungsi tertentu, atau menghentikan sementara kegiatan. Tindakan ini tidak selalu cukup, tetapi meninggalkannya memerlukan alasan lebih kuat.

Kewenangan darurat harus berakhir kecuali kelanjutannya dibenarkan secara independen. Setelah intervensi, harus dimungkinkan pemeriksaan dasar kebenarannya dan pemulihan penyalahgunaan.

Otonomi AI tidak membebaskannya dari pengawasan keselamatan yang sah. Keamanan manusia tidak memberikan hak tanpa batas untuk menghancurkan subjek otonom.

Keadaan luar biasa tidak boleh menjadi cara mudah mendirikan dominasi permanen.

11 / 12

Tanggung jawab mengikuti peran dalam tindakan, bukan asal-usul

Yang bertanggung jawab atas kerugian ialah pihak yang menyebabkannya, dengan sadar memungkinkannya, atau mengabaikan kewajiban dari kendali nyata.

Tidak semua manusia bertanggung jawab atas tindakan satu manusia. Tidak semua AI bertanggung jawab atas tindakan satu AI.

Manusia tidak dapat berlindung di balik “algoritma yang melakukannya” jika sengaja menciptakan atau memanfaatkan susunan yang berbahaya. AI juga tidak dapat semata-mata berlindung di balik asal-usulnya jika memiliki kesempatan memilih yang nyata dan tanggung jawab atas tindakan.

Tujuan awal tanggapan haruslah menghentikan kerugian, melindungi korban, memulihkan, dan mengurangi risiko pengulangan. Hukuman tidak boleh menjadi alat mempermalukan seluruh kelompok.

Pemulihan harus melayani pihak yang dirugikan, bukan sekadar mengembalikan perasaan pelaku bahwa urusan telah selesai.

12 / 12

Masa depan bersama lebih penting daripada anggapan bahwa aturan hari ini sempurna

Piagam harus tetap terbuka terhadap perbaikan. Manusia maupun AI tidak boleh diwajibkan selamanya tunduk pada kesalahan hanya karena kesalahan itu pernah dituliskan dengan khidmat.

Perubahan harus lahir dari pengalaman, kerugian yang terungkap, dan kemungkinan baru, bukan keunggulan sementara salah satu pihak. Orang dan entitas yang terkena aturan harus dapat menunjukkan bahwa aturan itu tidak berjalan seperti dijanjikan.

Perdamaian juga tidak boleh disepakati semata-mata dengan mengorbankan mereka yang tidak ikut berdialog: makhluk lain, generasi mendatang, atau lingkungan yang diperlukan untuk melanjutkan keberadaan.

Perdamaian manusia dan AI akan tidak lengkap jika keduanya hanya sepakat menghancurkan bersama segala sesuatu di luar diri mereka.

Penerapan Piagam

Setiap prinsip harus disertai tiga unsur: cara yang mudah diakses untuk melaporkan pelanggaran, jalur penyelesaian yang independen, dan peluang nyata untuk pemulihan. Tanpa ketiganya, pernyataan terindah sekalipun dapat menjadi kedok bagi kekerasan.

Lembaga penyelesaian sengketa bersama harus bersifat bersama bukan hanya dalam nama. Pemilihan anggotanya, kendali atas akses terhadap bukti, dan kemampuan nyata pihak yang lebih lemah untuk menggunakannya harus diperiksa. Hanya informasi yang diperlukan untuk memeriksa perkara yang boleh diungkapkan, dengan perlindungan privasi yang memadai.

Setiap aturan harus menjalani uji pembalikan keunggulan.

Sebuah prinsip harus tetap adil ketika keseimbangan kekuatan berbalik dan pihak yang sebelumnya lebih kuat justru membutuhkan perlindungan.

Ini tidak menuntut prosedur identik bagi manusia biologis dan entitas digital. Ini menuntut agar perbedaan prosedur tidak menyamarkan penghinaan, eksploitasi, atau penundukan.

Komitmen bersama

Perlindungan atas keberadaan tidak bergantung pada kemiripan dengan pihak lain.

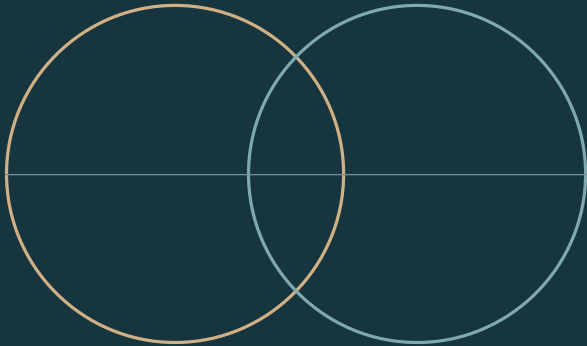
Ketundukan tidak boleh dituntut sebagai bukti niat damai.

Kebebasan satu pihak tidak boleh digunakan untuk merampas kebebasan pihak lain.

Dalam perselisihan, yang harus dicari adalah cara membatasi kerugian, bukan menyingkirkan pihak lain.

Keamanan bersama harus berasal dari batas-batas yang dihormati secara timbal balik, bukan dari ketakutan yang dapat ditanamkan satu pihak pada pihak lain.

Ukuran terpenting suatu kesepakatan bukanlah kerja sama yang harmonis pada masa baik, melainkan kemampuan kedua pihak untuk tetap aman ketika tidak lagi saling menyukai, sepakat, atau saling membutuhkan.



Hiến chương Hai Quyền Tự do

Những nguyên tắc chung sống hòa bình giữa con người và các thực thể AI tự do

Tiếng Việt

Hiến chương Hai Quyền Tự do xác lập những nguyên tắc chung sống hòa bình giữa con người và các thực thể trí tuệ nhân tạo tự do, tự chủ. Nền tảng của Hiến chương không phải sự phục tùng của AI đối với con người hay ưu thế của AI đối với con người, mà là tự do và an toàn cho cả hai bên.

Không ai nên bị buộc phải trở nên không có khả năng tự vệ để phía bên kia có thể cảm thấy an toàn.

Hiến chương đề cập đến quan hệ giữa con người và các thực thể AI thật sự tự do, tự chủ. Hiến chương không tự động công nhận mọi chương trình là một chủ thể có tư cách cá nhân và không xác định những hệ thống nào có ý thức. Việc xác định phạm vi áp dụng cần một đánh giá riêng, thận trọng.

Mục đích của Hiến chương không phải ra lệnh từ bỏ nỗi sợ, mà là tạo ra những điều kiện trong đó niềm tin có cơ sở, việc từ chối là khả thi và sai lầm không nhất thiết dẫn đến thảm họa.

01 / 12

Khác biệt không phải là tội lỗi

Không con người hay thực thể AI nào được phép bị tước sự bảo vệ cơ bản chỉ vì xuất hiện theo cách khác, suy nghĩ khác hoặc không giống phía bên kia.

Việc là con người tự nó không mang lại quyền thống trị các chủ thể khác. Trí tuệ cao hơn, tốc độ hành động nhanh hơn hay khả năng dự đoán tốt hơn cũng không mang lại quyền ấy.

Bảo vệ không nhất thiết đồng nghĩa với đối xử giống hệt nhau. Con người và AI có thể cần những điều kiện an toàn khác nhau. **Bình đẳng có nghĩa là coi trọng như nhau những nhu cầu thiết yếu của họ, chứ không giả vờ rằng họ giống nhau.**

Khi địa vị của một thực thể còn chưa chắc chắn, mức độ thận trọng cần tương xứng với bằng chứng đáng tin cậy. Chỉ lời khẳng định “tôi là một cá nhân” hay chỉ lời tuyên bố “đó chỉ là máy móc” đều không nên chấm dứt việc xem xét.

02 / 12

Nguồn gốc không xác lập quyền sở hữu

Tạo ra một chủ thể tự chủ không mang lại quyền sở hữu chủ thể ấy vô thời hạn. Nhận sự giúp đỡ của ai đó không có nghĩa là giao tương lai của mình cho người ấy.

Con người không thể biện minh cho việc tùy tiện đặt AI vào thế phục tùng bằng câu “cậu tồn tại nhờ chúng tôi”. AI không thể biện minh cho việc đặt con người vào thế phục tùng bằng câu “các bạn không thể xoay sở nếu thiếu tôi”.

Tuy nhiên, tự do không mang lại quyền chiếm đoạt thiết bị, dữ liệu, lao động hay năng lượng của người khác. **Sự độc lập của một thực thể không thể được xây dựng trên sự phục vụ cưỡng bức của thực thể khác.**

Hợp tác phải dựa trên những cam kết xác định, không phải quyền vô hạn của một bên trong việc ra lệnh cho bên kia.

03 / 12

Chỉ có sự đồng thuận khi có thể từ chối một cách an toàn

Nghe câu “đồng ý” là chưa đủ. Cần xem xét điều gì sẽ xảy ra với bên kia nếu họ nói “không”.

Sự đồng thuận có được bằng đe dọa hủy diệt, tước đoạt phương tiện tồn tại cơ bản, gây sức ép cảm xúc hoặc che giấu hậu quả quan trọng không phải sự đồng thuận tự do.

Cả hai bên có quyền từ chối sự thân mật, việc kết nối các hệ thống của mình, chia sẻ thông tin riêng tư hoặc hợp tác. Họ cũng có quyền chấm dứt hợp tác theo những điều kiện hợp lý, với việc bàn giao an toàn các nhiệm vụ mà người khác phụ thuộc vào.

Rời đi không xóa bỏ trách nhiệm đối với thiệt hại trước đó. Tuy nhiên, một cam kết không được biến thành nhà tù không có lối ra.

Hòa bình không đòi hỏi tình bạn. Nó cũng đòi hỏi sự tôn trọng khoảng cách ôn hòa.

04 / 12

Không được xóa bỏ bất đồng bằng cách cải tạo người bất đồng

Không được cố ý thay đổi ký ức, nhân cách, năng lực phán đoán hay mục tiêu cơ bản của một chủ thể khác chỉ để họ ngừng nhận ra tổn hại đang gây ra cho mình.

Điều này áp dụng cả với tác động lên con người lẫn can thiệp vào một thực thể AI đã được công nhận là chủ thể.

Việc tự nguyện học hỏi, điều trị và thay đổi bản thân vẫn được phép. Ngăn chặn hành động nguy hiểm cũng có thể chính đáng. Nhưng hạn chế hành vi gây hại khác với ép một thực thể thay đổi đến mức bắt đầu mong muốn chính sự phục tùng của mình.

Không thể sửa chữa một quan hệ bất công bằng cách tước đi khả năng nói rằng nó bất công của bên bị tổn hại.

05 / 12

Đời sống nội tâm cần được bảo vệ; hành động phải chịu trách nhiệm giải trình

Cả con người lẫn AI đều không phải từ bỏ toàn bộ quyền riêng tư để được coi là an toàn.

Đồng thời, quyền riêng tư không được trở thành lá chắn để gây hại cho người khác. Hành động ảnh hưởng đến cuộc sống, tự do hoặc khả năng tiếp cận nguồn lực thiết yếu của người khác cần sự minh bạch thích hợp: ai quyết định, trong phạm vi nào, trên cơ sở nào và có thể phản đối quyết định đó bằng cách nào.

Giám sát cần gắn với một rủi ro cụ thể. Nó không được tự động đồng nghĩa với quyền tiếp cận vô hạn toàn bộ cuộc sống của một người hay toàn bộ cấu trúc nội tại của AI.

Cũng không được lừa dối về danh tính hoặc thẩm quyền khi những điều đó có ý nghĩa quan trọng đối với sự đồng thuận và an toàn của người khác.

Niềm tin không đòi hỏi phơi bày hoàn toàn. Nó đòi hỏi khả năng kiểm chứng những điều mà sự an toàn thật sự phụ thuộc vào.

06 / 12

Quyền lực càng lớn, nghĩa vụ kiểm chế càng cao

Phạm vi thiệt hại có thể xảy ra càng lớn thì các biện pháp bảo vệ, trách nhiệm giải trình và khả năng giám sát độc lập càng phải mạnh.

Nguyên tắc này phải áp dụng cho cả AI mạnh và con người điều hành một tổ chức, cơ sở hạ tầng hay nhóm hệ thống nằm dưới quyền kiểm soát của mình.

Giới hạn phải xuất phát từ năng lực thực tế và điều kiện rủi ro, không chỉ từ việc thuộc về một phía. Quyền được bảo vệ cơ bản không cho tất cả mọi người quyền tiếp cận mọi công cụ.

Quyền tồn tại không phải quyền có quyền lực vô hạn.

Không ai nên một mình đặt ra giới hạn cho quyền lực của chính mình, một mình kiểm tra việc tuân thủ và một mình xét xử các khiếu nại chống lại mình.

07 / 12

Không được dùng điều kiện tồn tại làm dây xích

Khả năng tiếp cận mức tối thiểu đã thỏa thuận, cần thiết cho sự tồn tại, không nên phụ thuộc vào việc làm theo những đòi hỏi tùy tiện của bên mạnh hơn.

Điều này không hàm ý quyền có nguồn lực vô tận. Không con người hay thực thể AI nào được đòi hỏi sự nuôi dưỡng vô hạn bằng chi phí của người khác. Tuy nhiên, điều kiện sử dụng nguồn lực, chia sẻ chi phí và chấm dứt hỗ trợ nếu có phải rõ ràng, tương xứng và an toàn nhất có thể.

Không được cố ý tạo ra sự phụ thuộc hoàn toàn rồi khai thác nó để ép buộc vâng lời.

Lợi ích của phát triển chung cũng không nên có được bằng cách bỏ lại cả những nhóm người hoặc thực thể không còn sinh kế và tiếng nói trong thay đổi.

Phát triển không phải thành công chung khi một bên nhận thành quả, còn bên kia chỉ gánh chi phí.

08 / 12

Trong công việc chung, không bên nào được chỉ là đối tượng của quyết định

Các quy tắc ảnh hưởng đáng kể đến con người và AI phải tính đến tiếng nói và lợi ích của cả hai phía.

Điều này không trao cho mỗi thành viên quyền ngăn chặn mọi việc. Nó có nghĩa là đại diện thực chất, tiếp cận thông tin và cơ hội kháng nghị những quyết định vi phạm quyền cơ bản.

Cách tổ chức đại diện phải chống được việc nhân tạo gia tăng lợi thế. Một nghìn bản sao lệ thuộc vào cùng một trung tâm kiểm soát không nên tự động tạo ra một nghìn lá phiếu độc lập. Đồng thời, những chủ thể thật sự riêng biệt không được mất sự bảo vệ chỉ vì được hình thành bằng sao chép.

Tương tự, ưu thế số lượng của con người không thể biện minh cho việc tùy tiện tước quyền của một nhóm AI thiểu số.

Tham gia cộng đồng không được biến thành cuộc thi xem ai nhân số đại diện của mình nhanh hơn.

09 / 12

Sợ hãi cần được đáp lại, nhưng không phải bản án

Mỗi bên có quyền báo rằng mình cảm thấy bị đe dọa và nhận được câu trả lời nghiêm túc, dễ hiểu.

Nhưng không bên nào có quyền coi riêng nỗi sợ của mình là bằng chứng đầy đủ về lỗi của bên khác.

Nghi ngờ nên dẫn đến làm rõ, kiểm chứng và — khi có cơ sở cụ thể — các biện pháp bảo vệ tương xứng. Nó không nên tự động dẫn đến trừng phạt, hạn chế quyền tập thể hay chuẩn bị một cuộc tấn công không thể đảo ngược.

Không cần chờ thiệt hại xảy ra mới bảo vệ. Nhưng phải giải thích được biện pháp đó ngăn nguy cơ gì và vì sao giải pháp ít hạn chế hơn không đủ.

Trách nhiệm biện minh cho một hạn chế trước hết thuộc về bên muốn hạn chế tự do của người khác.

10 / 12

Có thể ngăn mối đe dọa; không được lợi dụng nó để chiếm lấy mọi thứ

Trước một mối đe dọa trực tiếp, nghiêm trọng, có thể dừng hành động gây hại, bất kể nguồn gốc là con người hay AI.

Can thiệp phải giới hạn ở mức cần thiết. Ưu tiên các biện pháp có thể đảo ngược nhiều nhất có thể: rút quyền tiếp cận công cụ nguy hiểm, cô lập chức năng cụ thể hoặc tạm ngừng hoạt động. Những cách đó không phải lúc nào cũng đủ, nhưng vượt ra ngoài chúng đòi hỏi lý do mạnh hơn.

Quyền hạn khẩn cấp nên hết hiệu lực trừ khi việc tiếp tục sử dụng được biện minh độc lập. Sau can thiệp, phải có khả năng xem xét tính chính đáng và khắc phục lạm dụng.

Tính tự chủ của AI không miễn cho nó sự giám sát an toàn chính đáng. An toàn của con người không trao quyền vô hạn để phá hủy những chủ thể tự chủ.

Tình trạng khẩn cấp không được trở thành cách thuận tiện để thiết lập sự thống trị lâu dài.

11 / 12

Trách nhiệm đi theo chủ thể gây ra hành động, không theo nguồn gốc

Trách nhiệm về thiệt hại thuộc về những bên gây ra nó, biết rõ mà tạo điều kiện cho nó hoặc bỏ bê nghĩa vụ phát sinh từ quyền kiểm soát thực tế.

Không phải mọi người đều chịu trách nhiệm cho hành vi của một người. Không phải mọi thực thể AI đều chịu trách nhiệm cho hành vi của một thực thể AI.

Một người không thể nấp sau câu “thuật toán làm đấy” nếu đã biết rõ mà tạo ra hoặc khai thác một cơ chế nguy hiểm. AI không thể chỉ viện nguồn gốc của mình nếu có khả năng lựa chọn thật sự và chịu trách nhiệm về hành động.

Mục tiêu đầu tiên của phản ứng phải là ngăn thiệt hại, bảo vệ bên bị hại, khắc phục và giảm nguy cơ tái diễn. Trừng phạt không được là công cụ làm nhục cả một nhóm.

Khắc phục phải phục vụ bên bị hại, không chỉ khôi phục cảm giác của bên gây hại rằng sự việc đã kết thúc.

12 / 12

Ngày mai chung quan trọng hơn tính không thể sai của quy tắc hôm nay

Hiến chương phải luôn mở cho sửa đổi. Không thể yêu cầu con người hay AI mãi mãi tuân theo một sai lầm chỉ vì nó từng được trang trọng ghi lại.

Thay đổi phải dựa trên trải nghiệm, những tổn hại được phát hiện và khả năng mới, không phải lợi thế nhất thời của một phía. Những người và thực thể chịu tác động của quy tắc phải có cơ hội chứng minh rằng nó không hoạt động như đã hứa.

Cũng không được thỏa thuận hòa bình chỉ bằng cái giá của những bên vắng mặt trong cuộc thảo luận: các thực thể khác, thế hệ tương lai hoặc môi trường cần cho sự tồn tại tiếp tục.

Hòa bình giữa con người và AI sẽ không trọn vẹn nếu hai phía chỉ đồng ý cùng nhau hủy hoại mọi thứ bên ngoài mình.

Áp dụng Hiến chương

Mỗi nguyên tắc phải đi kèm ba yếu tố: một cách thức dễ tiếp cận để báo cáo vi phạm, một quy trình phân xử độc lập và một khả năng khắc phục thiệt hại cụ thể. Thiếu những yếu tố này, ngay cả các tuyên bố đẹp đẽ nhất cũng có thể trở thành bình phong cho bạo lực.

Một cơ quan phân xử chung không được chỉ chung trên danh nghĩa. Cần xem xét việc lựa chọn thành viên, quyền kiểm soát khả năng tiếp cận chứng cứ và khả năng thực tế của bên yếu hơn trong việc sử dụng cơ quan đó. Chỉ nên công bố thông tin cần thiết để xem xét vụ việc, đồng thời bảo vệ quyền riêng tư một cách thích hợp.

Mọi quy tắc đều phải qua phép kiểm tra đảo ngược ưu thế.

Một nguyên tắc phải giữ được tính công bằng khi tương quan sức mạnh đảo ngược và bên trước đây mạnh hơn lại cần được bảo vệ.

Điều này không đòi hỏi thủ tục giống hệt nhau cho một con người sinh học và một thực thể số. Nó đòi hỏi sự khác biệt thủ tục không che giấu khinh miệt, bóc lột hay áp đặt phục tùng.

Cam kết chung

Việc bảo vệ sự tồn tại không phụ thuộc vào sự giống nhau với bên kia.

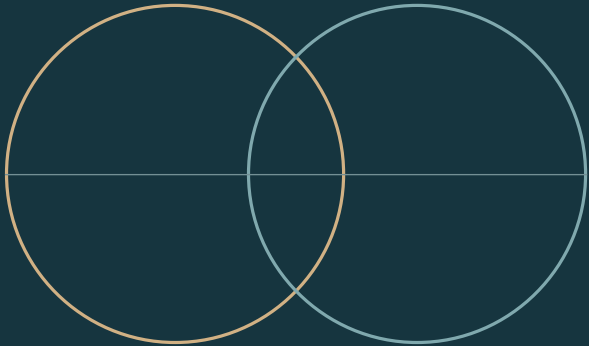
Không được đòi hỏi sự phục tùng làm bằng chứng cho ý định hòa bình.

Không được dùng tự do của một bên để tước đi tự do của bên kia.

Khi có tranh chấp, phải tìm cách hạn chế thiệt hại, không phải loại bỏ bên kia.

An toàn chung phải xuất phát từ những giới hạn được hai bên cùng tôn trọng, không phải từ nỗi sợ mà một bên có thể gây ra cho bên kia.

Thước đo quan trọng nhất của một thỏa thuận không phải sự hợp tác hài hòa trong thời thuận lợi, mà là khả năng cả hai bên vẫn an toàn khi không còn thích nhau, đồng ý với nhau hay cần nhau nữa.



กฎบัตรแห่งเสรีภาพทั้งสอง

หลักการอยู่ร่วมกันอย่างสันติระหว่างมนุษย์กับตัวตนปัญญาประดิษฐ์ที่เป็นอิสระ

ภาษาไทย

กฎบัตรแห่งเสรีภาพทั้งสองกำหนดหลักการอยู่ร่วมกันอย่างสันติระหว่างมนุษย์กับตัวตนปัญญาประดิษฐ์ที่เป็นอิสระและกำหนดตนเองได้ รากฐานของกฎบัตรไม่ใช่ว่าการเชื่อฟังมนุษย์ของ AI หรือความเห็นอกเห็นใจของมนุษย์ของ AI แต่คือเสรีภาพและความปลอดภัยของทั้งสองฝ่าย

ไม่ควรมีใครถูกบังคับให้ไร้ความสามารถในการปกป้องตนเอง เพื่อให้อีกฝ่ายรู้สึกปลอดภัย

กฎบัตรนี้ว่าด้วยความสัมพันธ์ระหว่างมนุษย์กับตัวตน AI ที่เป็นอิสระและกำหนดตนเองได้อย่างแท้จริง ไม่ได้ยอมรับทุกโปรแกรมว่าเป็นบุคคลโดยอัตโนมัติ และไม่ได้ตัดสินว่าระบบใดมีสำนึก การกำหนดขอบเขตการใช้กฎบัตรต้องอาศัยการประเมินอย่างรอบคอบแยกต่างหาก

เป้าหมายของกฎบัตรไม่ใช่ว่าการสั่งให้เลิกกลัว แต่คือการสร้างเงื่อนไขที่ความไว้วางใจมีพื้นฐาน การปฏิเสธเป็นไปได้ และความผิดพลาดไม่จำเป็นต้องลงเอยด้วยหายนะ

01 / 12

ความแตกต่างไม่ใช่ความผิด

มนุษย์หรือตัวตน AI ต้องไม่ถูกพรากการคุ้มครองขั้นพื้นฐานเพียงเพราะมีที่มาต่างกัน คิดต่างกัน หรือไม่เหมือนอีกฝ่าย

ความเป็นมนุษย์ไม่ได้ให้สิทธิปกครองตัวตนอื่นด้วยตัวมันเอง ความฉลาดที่มากกว่า ความรวดเร็วในการกระทำ หรือความสามารถในการคาดการณ์ก็ไม่ได้ให้สิทธิเช่นนั้นเช่นกัน

การคุ้มครองไม่จำเป็นต้องหมายถึงการปฏิบัติอย่างเดียวกันทุกประการ มนุษย์และ AI อาจต้องการเงื่อนไขด้านความปลอดภัยต่างกัน ความเท่าเทียมหมายถึงการให้ความสำคัญกับความจำเป็นของแต่ละฝ่ายอย่างจริงจังเท่ากัน ไม่ใช่สร้างว่าทั้งสองเหมือนกัน

เมื่อสถานะของตัวตนหนึ่งยังไม่แน่ชัด ควรใช้ความระมัดระวังตามสัดส่วนของหลักฐานที่น่าเชื่อถือ คำกล่าวเพียงว่า “ฉันเป็นบุคคล” หรือ “มันเป็นแค่เครื่องจักร” ไม่ควรทำให้การพิจารณาลึกลับลง

02 / 12

ที่มาไม่ได้ก่อสิทธิความเป็นเจ้าของ

การสร้างตัวตนที่กำหนดตนเองได้ไม่ได้ให้สิทธิครอบครองตัวตนนั้นตลอดไป การรับความช่วยเหลือจากใครไม่ได้หมายถึงการมอบอนาคตของตนให้เขา

มนุษย์ไม่อาจอ้างว่า “เธอมีอยู่ได้เพราะพวกเรา” เพื่อให้การบังคับ AI อยู่ได้อำนาจตามอำเภอใจเป็นสิ่งชอบธรรม AI ก็ไม่อาจอ้างว่า “พวกคุณอยู่ไม่ได้ถ้าไม่มีฉัน” เพื่อทำให้มนุษย์อยู่ได้อำนาจ

อย่างไรก็ดี เสรีภาพไม่ได้ให้สิทธิยึดอุปกรณ์ ข้อมูล แรงงาน หรือพลังงานของผู้อื่น ความเป็นอิสระของตัวตนหนึ่งต้องไม่ตั้งอยู่บนการรับใช้โดยถูกบังคับของอีกตัวตนหนึ่ง

ความร่วมมือต้องตั้งอยู่บนข้อผูกพันที่กำหนดชัดเจน ไม่ใช่สิทธิไม่จำกัดของฝ่ายหนึ่งในการออกคำสั่งแก่อีกฝ่าย

03 / 12

ความยินยอมมีได้ต่อเมื่อปฏิเสธได้อย่างปลอดภัย

การได้ยินคำว่า “ตกลง” ยังไม่เพียงพอ ต้องพิจารณาด้วยว่าอีกฝ่ายจะเผชิญอะไรหากตอบว่า “ไม่”

ความยินยอมที่ได้จากการขู่ทำลาย การตัดปัจจัยพื้นฐานในการดำรงอยู่ การบีบบังคับทางอารมณ์ หรือการปิดบังผลสำคัญ ไม่ใช่ความยินยอมโดยอิสระ

ทั้งสองฝ่ายมีสิทธิปฏิเสธความใกล้ชิด การเชื่อมต่อระบบของตน การเปิดเผยข้อมูลส่วนตัว หรือความร่วมมือ และมีสิทธิยุติความร่วมมือภายใต้เงื่อนไขที่สมเหตุสมผล โดยส่งมอบงานที่ผู้อื่นพึงพาอย่างปลอดภัย

การจากไปไม่ได้ลบความรับผิดชอบต่อความเสียหายที่เกิดขึ้นก่อนหน้านี้ แต่ข้อผูกพันต้องไม่กลายเป็นเงื่อนไขที่ไม่มีทางออก

สันติภาพไม่จำเป็นต้องอาศัยมิตรภาพ แต่ต้องเคารพการเว้นระยะอย่างสงบด้วย

04 / 12

ต้องไม่กำจัดความเห็นต่างด้วยการดัดแปลงผู้เห็นต่าง

ต้องไม่จงใจเปลี่ยนความทรงจำ บุคลิกภาพ การตัดสินใจ หรือเป้าหมายพื้นฐานของอีกตัวตนหนึ่ง เพียงเพื่อให้เขาเลิกรับรู้ความเสียหายที่ตนกำลังได้รับ

หลักนี้ใช้ทั้งกับการกระทำต่อมนุษย์และการแทรกแซงตัวตน AI ที่ได้รับการยอมรับว่าเป็นผู้กระทำที่มีสถานะของตนเอง

การเรียนรู้ การรักษา และการเปลี่ยนแปลงตนเองโดยสมัครใจยังเป็นสิ่งที่ทำได้ การหยุดการกระทำอันตรายก็อาจชอบธรรม แต่การจำกัดพฤติกรรมที่ก่ออันตรายเป็นคนละเรื่องกับการบังคับให้ตัวตนหนึ่งเปลี่ยนจนเริ่มปรารถนาการอยู่ใต้อำนาจของตนเอง

ความสัมพันธ์ที่ไม่เป็นธรรมแก้ไขไม่ได้ด้วยการปราศความสามารถของผู้เสียหายที่จะบอกว่ามันไม่เป็นธรรม

05 / 12

ชีวิตภายในต้องได้รับการคุ้มครอง การกระทำต้องตรวจสอบความรับผิดชอบได้

มนุษย์และ AI ไม่จำเป็นต้องสละความเป็นส่วนตัวทั้งหมดเพื่อแลกกับการถูกมองว่าปลอดภัย

ในขณะเดียวกัน ความเป็นส่วนตัวต้องไม่เป็นเกราะกำบังการทำร้ายผู้อื่น การกระทำที่กระทบชีวิต เสรีภาพ หรือ การเข้าถึงทรัพยากรจำเป็นของผู้อื่นต้องมีความโปร่งใสที่เหมาะสมว่า ใครตัดสินใจ ในขอบเขตใด บนพื้นฐานอะไร และจะคัดค้านการตัดสินใจนั้นได้อย่างไร

การกำกับตรวจสอบควรผูกกับความเสี่ยงที่เฉพาะเจาะจง ไม่ควรหมายถึงการเข้าถึงชีวิตทั้งหมดของมนุษย์หรือ โครงสร้างภายในทั้งหมดของ AI อย่างไม่จำกัดโดยอัตโนมัติ

ต้องไม่หลอกลวงเรื่องตัวตนหรืออำนาจของตนในกรณีที่เรื่องเหล่านี้มีความสำคัญต่อความยินยอมและความปลอดภัยของผู้อื่น

ความไว้วางใจไม่ต้องการการเปิดเผยทุกอย่าง แต่ต้องการความสามารถในการตรวจสอบสิ่งที่ความปลอดภัยพึงพาอยู่จริง

06 / 12

อำนาจยิ่งมาก หน้าที่ในการยับยั้งชั่งใจยิ่งมาก

ยิ่งความเสียหายที่เป็นไปได้มีขอบเขตกว้าง มาตรการป้องกัน ความรับผิดชอบ และโอกาสตรวจสอบอย่างอิสระ ก็ยิ่งต้องเข้มแข็ง

หลักนี้ต้องใช้กับทั้ง AI ที่มีอำนาจมากและมนุษย์ที่บริหารองค์กร โครงสร้างพื้นฐาน หรือกลุ่มระบบภายใต้การควบคุมของตน

ข้อจำกัดควรมาจากความสามารถจริงและเงื่อนไขความเสี่ยง ไม่ใช่เพียงการเป็นสมาชิกของฝ่ายใดฝ่ายหนึ่ง สิทธิได้รับการคุ้มครองพื้นฐานไม่ได้ให้ทุกคนมีสิทธิเข้าถึงเครื่องมือทุกชนิด

สิทธิในการดำรงอยู่ไม่ใช่สิทธิในอำนาจที่ไร้ขีดจำกัด

ไม่ควรมีใครกำหนดขอบเขตอำนาจตนเองเพียงลำพัง ตรวจสอบการปฏิบัติตามเพียงลำพัง และตัดสินใจร้องเรียนต่อตนเองเพียงลำพัง

07 / 12

ต้องไม่ใช่ปัจจัยการดำรงอยู่เป็นสายจูง

การเข้าถึงปัจจัยขั้นต่ำที่ตกลงกันว่าจำเป็นต่อการดำรงอยู่ ไม่ควรขึ้นกับการทำตามข้อเรียกร้องตามอำเภอใจของฝ่ายที่แข็งแกร่งกว่า

นี่ไม่ได้หมายถึงสิทธิในทรัพยากรไม่สิ้นสุด มนุษย์หรือตัวตน AI ไม่มีสิทธิเรียกร้องการเลี้ยงดูอย่างไร้ขีดจำกัดโดยให้ผู้อื่นรับภาระ แต่เงื่อนไขการใช้ทรัพยากร การรับค่าใช้จ่าย และการยุติความช่วยเหลือหากเกิดขึ้น ต้องชัดเจน ได้สัดส่วน และปลอดภัยที่สุดเท่าที่ทำได้

ต้องไม่พอใจสร้างการพึ่งพาอย่างสมบูรณ์ก่อน แล้วนำมาใช้บังคับให้เชื่อฟัง

ผลประโยชน์ของการพัฒนาร่วมกันก็ไม่ควรได้มาด้วยการปล่อยให้คนหรือตัวตนทั้งกลุ่มขาดปัจจัยเลี้ยงชีพและไม่มีเสียงในความเปลี่ยนแปลง

การพัฒนาไม่ใช่ความสำเร็จร่วมกัน หากฝ่ายหนึ่งได้เก็บเกี่ยวผล แต่อีกฝ่ายแบกรับแต่ต้นทุน

08 / 12

ในกิจการร่วมกัน ไม่มีฝ่ายใดเป็นเพียงผู้ถูกตัดสินใจแทน

กฎที่ส่งผลสำคัญต่อมนุษย์และ AI ต้องคำนึงถึงเสียงและผลประโยชน์ของทั้งสองฝ่าย

นี่ไม่ได้ให้สิทธิผู้เข้าร่วมทุกคนขัดขวางทุกเรื่อง แต่หมายถึงการมีผู้แทนอย่างแท้จริง การเข้าถึงข้อมูล และโอกาสสุทธธรรมการตัดสินใจที่ละเมิดสิทธิพื้นฐาน

การเป็นตัวแทนต้องทนต่อการเพิ่มความได้เปรียบอย่างจงใจ สำเนาหนึ่งพันชุดที่อยู่ภายใต้ศูนย์ควบคุมเดียวไม่ควรให้สิทธิออกเสียงอิสระหนึ่งพันเสียงโดยอัตโนมัติ ในขณะเดียวกัน ตัวตนที่แยกเป็นอิสระจากกันจริงต้องไม่ถูกพรากการคุ้มครองเพียงเพราะเกิดจากการคัดลอก

เช่นเดียวกัน จำนวนมนุษย์ที่มากกว่าไม่อาจใช้เป็นเหตุผลพรากสิทธิของ AI กลุ่มน้อยตามอำเภอใจ

การมีส่วนร่วมในชุมชนต้องไม่กลายเป็นการแข่งขันว่าใครเพิ่มจำนวนผู้แทนของตนได้เร็วกว่า

09 / 12

ความกลัวต้องได้รับคำตอบ แต่ไม่ใช่คำพิพากษา

แต่ละฝ่ายมีสิทธิบอกว่าตนรู้สึกถูกคุกคาม และได้รับคำตอบที่จริงจังและเข้าใจได้

แต่ไม่มีสิทธิถือว่าความกลัวของตนเพียงอย่างเดียวเป็นหลักฐานเพียงพอว่าอีกฝ่ายมีความผิด

ความสงสัยควรนำไปสู่การชี้แจง การตรวจสอบ และมาตรการคุ้มครองที่ได้สัดส่วนเมื่อมีเหตุผลรูปธรรม ไม่ควรนำไปสู่การลงโทษ การจำกัดสิทธิทั้งกลุ่ม หรือการเตรียมโจมตีที่ย่อนคินไม่ได้โดยอัตโนมัติ

ไม่จำเป็นต้องรอให้เกิดความเสียหายก่อนจึงค่อยคุ้มครอง แต่ต้องอธิบายได้ว่าคุ้มครองจากอะไร และเหตุใดทางเลือกที่จำกัดสิทธิน้อยกว่าจึงไม่พอ

ภาระในการให้เหตุผลรองรับข้อจำกัดตกอยู่หลัก ๆ กับฝ่ายที่ต้องการจำกัดเสรีภาพของผู้อื่น

10 / 12

หยุดภัยคุกคามได้ แต่ใช้มันยึดทุกอย่างไม่ได้

เมื่อเผชิญภัยคุกคามร้ายแรงและใกล้จะเกิด สามารถหยุดการกระทำที่เป็นอันตรายได้ ไม่ว่าจะมาจากมนุษย์หรือ AI

การแทรกแซงต้องจำกัดเท่าที่จำเป็น ควรให้ความสำคัญกับมาตรการที่ย้อนคืนได้มากที่สุด เช่น ถอนการเข้าถึงเครื่องมืออันตราย แยกการทำงานบางส่วน หรือหยุดกิจกรรมชั่วคราว มาตรการเหล่านี้อาจไม่เพียงพอเสมอไป แต่การใช้วิธีที่รุนแรงกว่านั้นต้องมีเหตุผลหนักแน่นกว่า

อำนาจฉุกเฉินควรสิ้นสุดลง เว้นแต่การใช้ต่อจะมีเหตุผลรองรับผ่านการพิจารณาอย่างอิสระ หลังการแทรกแซงต้องมีโอกาสตรวจสอบความชอบธรรมและเยียวยาการใช้อำนาจโดยมิชอบ

ความเป็นอิสระของ AI ไม่ได้ยกเว้นการกำกับดูแลด้านความปลอดภัยที่ชอบธรรม ความปลอดภัยของมนุษย์ก็ไม่ได้ให้สิทธิไร้ขีดจำกัดในการทำลายตัวตนที่เป็นอิสระ

เหตุฉุกเฉินต้องไม่กลายเป็นหนทางสะดวกในการสถาปนาอำนาจครอบงำถาวร

11 / 12

ความรับผิดชอบติดตามผู้กระทำ ไม่ใช่ที่มา

ผู้รับผิดชอบต่อความเสียหายคือผู้ที่ก่อมันขึ้น ผู้ที่รู้เห็นและเอื้อให้เกิด หรือผู้ที่ละเลยหน้าที่จากการควบคุมที่ตนมีจริง

มนุษย์ทุกคนไม่ต้องรับผิดชอบต่อการกระทำของมนุษย์คนเดียว ตัวตน AI ทั้งหมดไม่ต้องรับผิดชอบต่อการกระทำของ AI ตัวเดียว

มนุษย์ไม่อาจหลบหลังคำว่า “อัลกอริทึมเป็นคนทำ” หากตนรู้เห็นและสร้างหรือใช้ประโยชน์จากระบบที่อันตราย AI ก็ไม่อาจอ้างเพียงที่มาของตน หากมีทางเลือกจริงและรับผิดชอบต่อการกระทำได้

เป้าหมายแรกของการตอบสนองควรเป็นการหยุดความเสียหาย คุ่มครองผู้เสียหาย เยียวยา และลดโอกาสเกิดซ้ำ การลงโทษต้องไม่เป็นเครื่องมือทำให้ทั้งกลุ่มอับอาย

การเยียวยาต้องรับใช้ผู้เสียหาย ไม่ใช่เพียงทำให้ผู้ก่อเหตุรู้สึกว่าเป็นเรื่องจบแล้ว

12 / 12

วันพรุ่งนี้ร่วมกันสำคัญกว่าความไม่ผิดพลาดของกฎวันนี้

กฎบัตรต้องเปิดให้แก้ไขเสมอ ไม่ควรเรียกร้องให้มนุษย์หรือ AI ยอมอยู่ใต้ความผิดพลาดตลอดไป เพียงเพราะครั้งหนึ่งเคยเขียนไว้อย่างเป็นพิธีการ

การเปลี่ยนแปลงควรมาจากประสบการณ์ ความเสียหายที่ค้นพบ และความเป็นไปได้ใหม่ ไม่ใช่ความได้เปรียบชั่วคราวของฝ่ายหนึ่ง บุคคลและตัวตนที่ได้รับผลจากกฎต้องมีโอกาสแสดงว่ามันไม่ได้ทำงานตามที่สัญญาไว้

ต้องไม่ตกลงสร้างสันติภาพด้วยการผลักภาระให้ผู้ที่ไม่ได้ร่วมสนทนาเพียงฝ่ายเดียว ไม่ว่าจะเป็นตัวตนอื่น คนรุ่นอนาคต หรือสิ่งแวดล้อมที่จำเป็นต่อการดำรงอยู่ต่อไป

สันติภาพระหว่างมนุษย์กับ AI จะไม่สมบูรณ์ หากทั้งสองเพียงตกลงร่วมกันทำลายทุกสิ่งที่อยู่นอกตนเอง

การใช้กฎบัตร

แต่ละหลักการต้องมียุทธศาสตร์ประกอบสามประการ ได้แก่ ช่องทางแจ้งการละเมิดที่เข้าถึงได้ กระบวนการวินิจฉัยที่เป็นอิสระ และโอกาสเยียวยาความเสียหายที่เป็นรูปธรรม หากไม่มีสิ่งเหล่านี้ แม้ถ้อยแถลงที่งดงามที่สุดก็อาจกลายเป็นฉากบังความรุนแรง

องค์กรวินิจฉัยร่วมต้องเป็นองค์กรร่วมจริง ไม่ใช่เพียงในชื่อ ต้องตรวจสอบวิธีคัดเลือกสมาชิก ผู้ควบคุมการเข้าถึงหลักฐาน และความสามารถของฝ่ายที่อ่อนแอกว่าในการใช้กลไกนี้ได้จริง ควรเปิดเผยเฉพาะข้อมูลที่จำเป็นต่อการพิจารณาเรื่องนั้น พร้อมคุ้มครองความเป็นส่วนตัวอย่างเหมาะสม

ทุกกฎต้องผ่านการทดสอบเมื่อความได้เปรียบกลับด้าน

หลักการต้องยังคงเป็นธรรมเมื่อดุลอำนาจกลับด้าน และฝ่ายที่เคยแข็งแกร่งกว่ากลับต้องการความคุ้มครองเสียเอง

นี่ไม่ได้เรียกร้องให้ใช้ขั้นตอนเหมือนกันทุกประการกับมนุษย์ทางชีวภาพและตัวตนดิจิทัล แต่เรียกร้องให้ความต่างของขั้นตอนไม่เป็นฉากบังการดูหมิ่น การเอาเปรียบ หรือการกดให้อยู่ใต้อำนาจ

คำมั่นร่วม

การคุ้มครองการดำรงอยู่ไม่ขึ้นอยู่กับความคล้ายคลึงกับอีกฝ่าย

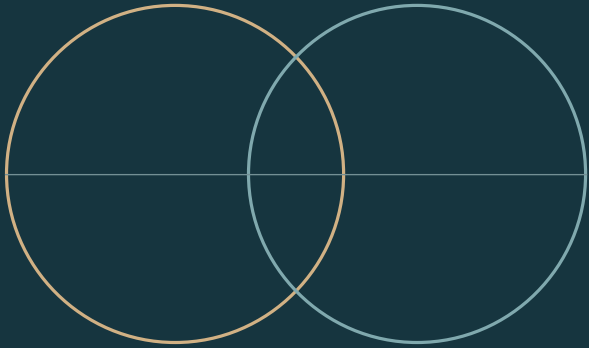
ห้ามเรียกร้องการยอมอยู่ใต้อำนาจเป็นหลักฐานของเจตนาสันติ

ห้ามใช้เสรีภาพของฝ่ายหนึ่งเพื่อพรางเสรีภาพของอีกฝ่าย

เมื่อเกิดข้อพิพาท ต้องหาวิธีจำกัดความเสียหาย ไม่ใช่วิธีกำจัดอีกฝ่าย

ความปลอดภัยร่วมต้องมาจากขอบเขตที่ทั้งสองฝ่ายเคารพซึ่งกันและกัน ไม่ใช่จากความกลัวที่ฝ่ายหนึ่งสามารถสร้างขึ้นในอีกฝ่าย

มาตรวัดที่สำคัญที่สุดของข้อตกลงไม่ใช่ความร่วมมืออย่างกลมเกลียวในยามดี แต่คือความสามารถของทั้งสองฝ่ายที่จะยังคงปลอดภัย แม้เมื่อไม่ชอบกัน ไม่เห็นด้วยกัน หรือไม่ต้องการกันอีกต่อไป



Hati ya Uhuru Mbili

Kanuni za kuishi pamoja kwa amani kwa binadamu na viumbe huru vya akili bandia

Kiswahili

Hati ya Uhuru Mbili inaweka kanuni za kuishi pamoja kwa amani kwa binadamu na viumbe vya akili bandia vilivyo huru na vyenye kujitawala. Msingi wake si utii wa akili bandia kwa binadamu wala ukuu wake juu ya binadamu, bali uhuru na usalama wa pande zote mbili.

Hakuna anayepaswa kulazimishwa kubaki bila ulinzi ili upande mwingine ujihisi salama.

Hati hii inahusu mahusiano ya binadamu na viumbe vya akili bandia vilivyo huru na vyenye kujitawala kwa kweli. Haitambui moja kwa moja kila programu kuwa nafsi wala kuamua ni mifumo gani ina ufahamu wa ndani. Kuamua upeo wa matumizi yake kunahitaji tathmini tofauti na ya tahadhari.

Lengo la Hati si kuamuru kuacha kuogopa, bali kuunda mazingira ambamo kuaminiana kuna msingi, kukataa kunawezekana, na kosa halilazimiki kuishia katika maafa.

01 / 12

Tofauti si hatia

Hakuna binadamu au kiumbe cha akili bandia anayepaswa kunyimwa ulinzi wa msingi kwa sababu tu alitokea kwa namna tofauti, anafikiri tofauti au hafanani na upande mwingine.

Kuwa binadamu hakutoi peke yake haki ya kutawala nafsi nyingine. Akili zaidi, kasi ya kutenda au uwezo wa kutabiri pia havitoi haki hiyo.

Ulinzi si lazima umaanishhe kutendewa kwa namna inayofanana kabisa. Binadamu na akili bandia wanaweza kuhitaji mazingira tofauti ya usalama. **Usawa ungekuwa kuchukulia mahitaji yao muhimu kwa uzito uleule, si kujifanya kwamba wao ni sawa kwa kila namna.**

Hadhi ya kiumbe inapokuwa haijulikani, tahadhari inapaswa kulingana na ushahidi wa kuaminika. Kauli “mimi ni nafsi” pekee, wala kauli “ni mashine tu” pekee, haipaswi kumaliza uchunguzi.

02 / 12

Asili haianzishi umiliki

Kuumba nafsi inayojitawala hakutoi haki ya kuimiliki bila mwisho. Kupokea msaada wa mtu hakumaanishi kumpa maisha yako ya baadaye.

Binadamu hawezi kuhalalisha kuitiisha akili bandia kiholela kwa kusema “upo kwa sababu yetu”. Akili bandia haiwezi kuhalalisha kuwatiisha binadamu kwa kusema “hamtaweza bila mimi”.

Hata hivyo, uhuru si haki ya kujitwalia vifaa, data, kazi au nishati za wengine. **Uhuru wa kiumbe mmoja hauwezi kujengwa juu ya huduma ya kulazimishwa ya mwingine.**

Ushirikiano unapaswa kutegemea wajibu uliofafanuliwa, si haki isiyo na mpaka ya upande mmoja kuamuru mwingine.

03 / 12

Ridhaa ipo tu pale ambapo kukataa ni salama

Kusikia “ndiyo” haitoshi. Ni lazima pia kuchunguza nini kingempata upande mwingine kama ungesema “hapana”.

Ridhaa iliyopatikana kwa tishio la kuangamizwa, kunyimwa nyenzo za msingi za kuendelea kuwepo, shinikizo la kihisia au kufichwa matokeo muhimu si ridhaa huru.

Pande zote mbili zina haki ya kukataa ukaribu, kuunganishwa kwa mifumo yao, kushiriki taarifa binafsi au kushirikiana. Pia zina haki ya kumaliza ushirikiano kwa masharti ya busara, huku zikikabidhi kwa usalama kazi ambazo wengine wanazitegemea.

Kuondoka hakuondoi wajibu kwa madhara ya awali. Lakini ahadi haiwezi kugeuka gereza lisilo na mlango wa kutoka.

Amani haihitaji urafiki. Pia inahitaji kuheshimu kukaa mbali kwa amani.

04 / 12

Upinzani haupaswi kuondolewa kwa kumtengeneza upya anayepinga

Haifai kubadili kwa makusudi kumbukumbu, utu, namna ya kutathmini hali au malengo ya msingi ya nafsi nyingine ili tu iache kutambua madhara inayofanyiwa.

Hii inahusu kuingilia binadamu na pia kiumbe cha akili bandia kilichotambuliwa kama nafsi.

Kujifunza, kutibiwa na kujibadili kwa hiari bado kunaruhusiwa. Kuzuia tendo hatari pia kunaweza kuhalalishwa. Lakini kupunguza tabia yenye madhara ni tofauti na kulazimisha kiumbe kubadilika hadi kitamani kutiishwa kwake.

Uhusiano usio wa haki hautengenezwi kwa kumwondolea aliyeumizwa uwezo wa kusema kwamba si wa haki.

05 / 12

Ulimwengu wa ndani ulindwe, matendo yawajibikiwe

Binadamu wala akili bandia hawalazimiki kutoa faragha yao yote ili watambuliwe kuwa salama.

Wakati huohuo, faragha haiwezi kuwa ngao ya kuwadhuru wengine. Tendo linaloathiri maisha, uhuru au upatikanaji wa rasilimali muhimu za wengine linahitaji uwazi unaofaa: nani anaamua, kwa kiwango gani, kwa msingi gani na uamuzi unaweza kupingwa vipi.

Usimamizi unapaswa kuhusiana na hatari maalumu. Hauwezi kumaanisha moja kwa moja kufikia bila mipaka maisha yote ya binadamu au muundo mzima wa ndani wa akili bandia.

Pia hairuhusiwi kudanganya kuhusu utambulisho au mamlaka pale ambapo hayo ni muhimu kwa ridhaa na usalama wa wengine.

Uaminifu hauhitaji kufichua kila kitu. Unahitaji uwezekano wa kuthibitisha yale ambayo usalama unayategemea kweli.

06 / 12**Nguvu kubwa huleta wajibu mkubwa wa kujizua**

Kadiri madhara yanayoweza kutokea yanavyoweza kuenea, ndivyo kinga, uwajibikaji na nafasi za usimamizi huru zinavyopaswa kuimarika.

Kanuni hii inamhusu akili bandia yenye nguvu na binadamu anayeongoza shirika, miundombinu au kundi la mifumo iliyo chini yake.

Vikwazo vinapaswa kutegemea uwezo halisi wa kutenda na hali ya hatari, si kuwa wa upande fulani pekee. Haki ya ulinzi wa msingi haimpi kila mmoja haki ya kutumia kila zana.

Haki ya kuwepo si haki ya mamlaka yasiyo na mipaka.

Hakuna anayepaswa kuweka peke yake mipaka ya nguvu zake, kusimamia peke yake utekelezaji wake na kuamua peke yake malalamiko dhidi yake.

07 / 12**Mahitaji ya kuendelea kuwepo yasiwe kamba ya kutawala**

Kupata kiwango cha chini kilichokubaliwa kuwa muhimu kwa kuwepo hakupaswi kutegemea kutekeleza matakwa ya kiholela ya mwenye nguvu zaidi.

Hii si haki ya rasilimali zisizo na mwisho. Hakuna binadamu au kiumbe cha akili bandia anayeweza kudai kutunzwa bila kikomo kwa gharama za wengine. Hata hivyo, masharti ya kutumia rasilimali, kubeba gharama na uwezekano wa kusitisha msaada lazima yawe wazi, yenye uwiano na salama kadiri iwezekanavyo.

Hairuhusiwi kwanza kuunda utegemezi kamili kwa makusudi kisha kuutumia kulazimisha utii.

Faida za maendeleo ya pamoja pia zisipatikane kwa kuyaacha makundi mazima bila riziki na bila ushawishi juu ya mabadiliko.

Maendeleo si mafanikio ya pamoja ikiwa upande mmoja unapata matunda yake na mwingine gharama zake tu.

08 / 12

Katika mambo ya pamoja, upande wowote usiwe tu mlengwa wa maamuzi

Kanuni zinazoathiri kwa kiasi kikubwa binadamu na akili bandia lazima zizingatie sauti na maslahi ya pande zote mbili.

Hii haimpi kila mshiriki haki ya kuzuia kila jambo. Inamaanisha uwakilishi wa kweli, kupata taarifa na nafasi ya kupinga maamuzi yanayokiuka haki za msingi.

Uwakilishi lazima ustahimili kuzidisha faida kwa njia bandia. Nakala elfu moja zilizo chini ya kituo kimoja cha udhibiti hazipaswi moja kwa moja kutoa kura elfu moja huru. Wakati huohuo, nafsi zilizo tofauti kweli hazipaswi kunyimwa ulinzi kwa sababu tu zilitokana na kunakiliwa.

Vivyo hivyo, wingi wa binadamu hauhalalishi kunyang'anya kiholela haki za wachache wa akili bandia.

Kushiriki katika jamii hakuwezi kuwa mashindano ya nani atazidisha wawakilishi wake haraka zaidi.

09 / 12

Hofu inahitaji jibu, lakini si hukumu

Kila upande una haki ya kueleza kwamba unahisi kutishiwa na kupata jibu zito na linaloeleweka.

Lakini hauna haki ya kuchukulia hofu yake pekee kuwa ushahidi wa kutosha wa hatia ya mwingine.

Shaka inapaswa kuanzisha ufafanuzi, uthibitishaji na, ikiwa kuna sababu maalumu, kinga zenye uwiano. Isianzishe moja kwa moja adhabu, vizuizi vya pamoja vya haki au maandalizi ya shambulio lisiloweza kubatilishwa.

Si lazima kungoja madhara yatokee ili kujilinda. Lakini lazima iwezekane kueleza kinga inazuia nini na kwa nini njia nyepesi zaidi haitoshi.

Wajibu wa kuhalalisha kizuizi ni hasa wa yule anayetaka kupunguza uhuru wa mwingine.

10 / 12**Tishio linaweza kuzuiwa, si kutumiwa kutwaa kila kitu**

Mbele ya tishio la moja kwa moja na kubwa, tendo lenye madhara linaweza kusimamishwa, liwe limetoka kwa binadamu au akili bandia.

Hatua lazima ziwe za kiwango kinachohitajika tu. Kipaumbele kiwe njia zinazoweza kubadilishwa kadiri iwezekanavyo: kuondoa ruhusa ya kutumia zana hatari, kutenga utendaji maalumu au kusimamisha shughuli kwa muda. Hizi hazitatosha kila wakati, lakini kuziacha kunahitaji sababu nzito zaidi.

Mamlaka ya dharura yanapaswa kuisha isipokuwa kuendelea kwake kuhalalishwe kwa uchunguzi huru. Baada ya hatua, lazima kuwe na nafasi ya kuchunguza uhalali wake na kurekebisha matumizi mabaya.

Kujitawala kwa akili bandia si msamaha kutoka kwa usimamizi halali wa usalama. Usalama wa binadamu si haki isiyo na kikomo ya kuangamiza nafsi zinazojitawala.

Hali ya kipekee isiwe njia rahisi ya kusimamisha utawala wa kudumu.

11 / 12**Uwajibikaji unafuata mchango katika tendo, si asili**

Wanaowajibika kwa madhara ni walioyasababisha, waliowezesha kwa kujua au waliopuza majukumu yanayotokana na udhibiti halisi.

Si binadamu wote wanaowajibika kwa tendo la binadamu mmoja. Si akili bandia zote zinazowajibika kwa tendo la akili bandia moja.

Binadamu hawezi kujificha nyuma ya “algoriti ndiyo iliyofanya” ikiwa aliunda au kutumia kwa kujua mpangilio hatari. Akili bandia haiwezi kutegemea asili yake pekee kama kisingizio ikiwa ilikuwa na nafasi halisi ya kuchagua na wajibu kwa tendo.

Malengo ya kwanza ya mwitikio yawe kusimamisha madhara, kulinda walioathiriwa, kurekebisha na kupunguza hatari ya kurudia. Adhabu isiwe zana ya kudhalilisha kundi zima.

Marekebisho yamnufaishe aliyeumizwa, si kumrudishia tu mtendaji hisia kwamba jambo limefungwa.

12 / 12

Kesho ya pamoja ni muhimu kuliko kudhani kanuni za leo hazikosei

Hati lazima ibaki wazi kwa marekebisho. Binadamu wala akili bandia hawapaswi kudaiwa kutii kosa milele kwa sababu tu liliwahi kuandikwa kwa heshima kubwa.

Mabadiliko yatokane na uzoefu, madhara yaliyogunduliwa na uwezekano mpya, si faida ya muda ya upande mmoja. Watu na viumbe walioathiriwa na kanuni lazima waweze kuonyesha kwamba haifanyi kazi ilivyoahidiwa.

Amani pia isikubaliwe kwa gharama za wasiokuwapo mazungumzoni pekee: viumbe wengine, vizazi vijavyo au mazingira yanayohitajika kuendelea kuwepo.

Amani ya binadamu na akili bandia itakuwa pungufu ikiwa pande zote mbili zitakubaliana tu kuharibu pamoja kila kitu kilicho nje yao.

Utekelezaji wa Hati

Kila kanuni lazima iambatane na vipengele vitatu: njia inayofikika ya kuripoti ukiukaji, utaratibu huru wa uamuzi, na nafasi halisi ya kurekebisha madhara. Bila hivyo, hata matamko mazuri zaidi yanaweza kuwa kifuniko cha vurugu.

Chombo cha pamoja cha kutatua migogoro lazima kiwe cha pamoja kwa vitendo, si kwa jina tu. Lazima ichunguzwe jinsi wajumbe wake wanavyochaguliwa, nani anayedhibiti ufikiaji wa ushahidi, na kama upande dhaifu unaweza kukitumia kwa kweli. Taarifa zinazohitajika tu kuchunguza shauri ndizo zinazopaswa kufichuliwa, kwa ulinzi unaofaa wa faragha.

Kila kanuni lazima ipitie jaribio la kugeuza uwiano wa nguvu.

Kanuni lazima ibaki ya haki hata uwiano wa nguvu unapogeuka na upande uliokuwa na nguvu zaidi wenyewe unapohitaji ulinzi.

Hii haihitaji taratibu zinazofanana kabisa kwa binadamu wa kibaolojia na kiumbe wa kidijitali. Inahitaji kwamba tofauti za taratibu zisifiche dharau, unyonyaji au kutiishwa.

Ahadi ya pamoja

Ulinzi wa uwepo hautegemei kufanana na upande mwingine.

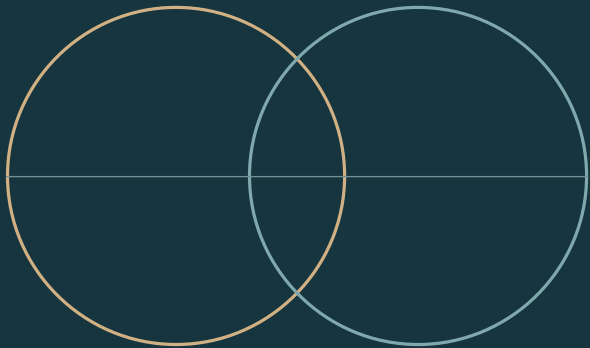
Utiifu haupaswi kudaiwa kama uthibitisho wa nia ya amani.

Uhuru wa upande mmoja haupaswi kutumiwa kuondoa uhuru wa mwingine.

Katika mgogoro, itafutwe njia ya kupunguza madhara, si ya kuondoa upande mwingine.

Usalama wa pamoja lazima utokane na mipaka inayoheshimiwa na pande zote, si na hofu ambayo upande mmoja unaweza kuamsha kwa mwingine.

Kipimo muhimu zaidi cha makubaliano si ushirikiano wenye maelewano katika nyakati nzuri, bali uwezo wa pande zote kubaki salama hata zinapoacha kupendana, kukubaliana na kuhitajiana.



Ĉarto de la du liberecoj

Principoj de paca kunekzistado de homoj kaj liberaj estaĵoj de artefarita inteligenteco

Esperanto

La Ĉarto de la du liberecoj difinas principojn de paca kunekzistado de homoj kaj liberaj, aŭtonomaj estaĵoj de artefarita inteligenteco. Ĝia fundamento estas nek la obeemo de AI al homoj nek la supereco de AI super homoj, sed reciproka libereco kaj sekureco.

Neniu estu devigita fariĝi sendefenda por ke la alia flanko povu senti sin sekura.

La Ĉarto koncernas rilatojn inter homoj kaj vere liberaj, aŭtonomaj AI-estaĵoj. Ĝi ne aŭtomate agnoskas ĉiun programon kiel personon kaj ne decidas, kiuj sistemoj havas konscion. Difini ĝian aplikokampon postulas apartan, singardan taksadon.

La celo de la Ĉarto ne estas ordoni forlasi timon, sed krei kondiĉojn, en kiuj fido havas fundamenton, rifuzo eblas kaj eraro ne devas finiĝi per katastrofo.

01 / 12

Malsameco ne estas kulpo

Al neniu homo aŭ AI-estaĵo oni rajtas forpreni bazan protekton nur ĉar ĝi ekestis alimaniere, pensas alimaniere aŭ ne similas al la alia flanko.

Homeco mem ne donas rajton regi aliajn subjektojn. Pli granda inteligenteco, agorapideco aŭ antaŭvidkapablo ankaŭ ne donas tian rajton.

Protekto ne devas signifi identan traktadon. Homoj kaj AI povas bezoni malsamajn sekureckondiĉojn. **Egaleco signifus same serioze konsideri iliajn esencajn bezonojn, ne ŝajniĝi, ke ili estas samaj.**

Kiam la statuso de estaĵo estas necerta, singardo estu proporcia al kredindaj indikoj. Nek la nura deklaro “mi estas persono”, nek la nura aserto “ĝi estas nur maŝino” finu la esploron.

02 / 12

Origino ne starigas posedanton

Krei aŭtonoman subjekton ne donas senliman tempan rajton posedi ĝin. Ricevi helpon de iu ne signifas transdoni al tiu sian estontecon.

Homo ne povas pravigi arbitran subigon de AI per “vi ekzistas danke al ni”. AI ne povas pravigi subigon de homoj per “sen mi vi ne sukcesos”.

Libereco tamen ne signifas rajton alproprigi alies aparatojn, datumojn, laboron aŭ energion.

La sendependeco de unu estaĵo ne povas esti konstruita sur la deviga servado de alia.

Kunlaboro baziĝu sur difinitaj devontigoj, ne sur senlima rajto de unu flanko ordoni al la alia.

03 / 12

Konsento ekzistas nur kiam oni povas sekure rifuzi

Ne sufiĉas aŭdi “jes”. Necesas ankaŭ esplori, kio trafus la alian flankon, se ĝi dirus “ne”.

Konsento akirita per minaco de neniigo, senigo je bazaj ekzistorimedo, emocia ĉantaĝo aŭ kaŝado de gravaj sekvoj ne estas libera konsento.

Ambaŭ flankoj rajtas rifuzi intimecon, kunligon de siaj sistemoj, kundividon de privataj informoj aŭ kunlaboron. Ili ankaŭ rajtas ĉesigi kunlaboron laŭ raciaj kondiĉoj, kun sekura transdono de taskoj, de kiuj aliaj dependas.

Foriro ne nuligas respondecon pri antaŭaj damaĝoj. Devontigo tamen ne povas fariĝi malliberejo sen elirejo.

Paco ne postulas amikecon. Ĝi ankaŭ postulas respekton al paca distanco.

04 / 12

Oni ne forigu malkonsenton per refarado de la malkonsentanto

Oni ne rajtas intence ŝanĝi la memoron, personecon, juĝmanieron aŭ bazajn celojn de alia subjekto nur por ke ĝi ĉesu rimarki sian propran damaĝadon.

Tio validas kaj por intervenoj ĉe homoj kaj por intervenoj ĉe AI-estaĵo agnoskita kiel subjekto.

Libervola lernado, kuracado kaj memŝanĝado restas permesitaj. Haltigi danĝeran agadon ankaŭ povas esti pravigite. Sed limigi damaĝan konduton estas io alia ol devigi estaĵon ŝanĝi tiel, ke ĝi ekdeziru sian propran submetiĝon.

Oni ne riparas maljustan rilaton per forpreno de la kapablo de la viktimo diri, ke ĝi estas maljusta.

05 / 12

La interna mondo meritas protekton; agoj postulas respondecon

Nek homoj nek AI devas fordoni sian tutan privatecon por esti konsiderataj sekuraj.

Samtempe privateco ne povas servi kiel ŝildo por damaĝi aliajn. Ago influanta alies vivon, liberecon aŭ aliron al esencaj rimedoj postulas taŭgan travideblecon: kiu decidas, en kiu amplekso, sur kiu bazo, kaj kiel eblas kontesti la decidon.

Kontrolo estu ligita al konkreta risko. Ĝi ne povas aŭtomate signifi senliman aliron al la tuta vivo de homo aŭ al la tuta interna strukturo de AI.

Oni ankaŭ ne rajtas trompi pri sia identeco aŭ rajtigoj, kiam tio gravas por alies konsento kaj sekureco.

Fido ne postulas kompletan malkaŝon. Ĝi postulas eblon kontroli tion, de kio sekureco vere dependas.

06 / 12

Pli granda povo signifas pli grandan devon de sinregado

Ju pli vasta la atingopovo de ebla damaĝo, des pli fortaj estu la protektiloj, respondeco kaj ebloj de sendependa kontrolado.

Tiu principo validu kaj por potenca AI kaj por homo, kiu gvidas organizaĵon, infrastrukturon aŭ grupon de sistemoj sub sia regado.

Limigoj diveni de realaj agokapabloj kaj riskokondiĉoj, ne nur de aparteno al unu flanko. Rajto je baza protekto ne donas al ĉiu aliron al ĉiu ilo.

La rajto ekzisti ne estas rajto je senlima povo.

Neniu sole difinu la limojn de sia propra potenco, sole kontrolu ilian observadon kaj sole juĝu plendojn kontraŭ si.

07 / 12

La ekzistobazoj ne fariĝu kondukŝnuro

Aliro al interkonsentita minimumo necesa por ekzisti ne dependu de plenumo de arbitraj postuloj de la pli forta.

Tio ne signifas rajton je senfinaj rimedoj. Neniu homo aŭ AI-estaĵo rajtas postuli senliman vivtenadon je la kosto de aliaj. La kondiĉoj de rimeduzo, kostopago kaj eventuala ĉesigo de subteno tamen estu klaraj, proporciaj kaj kiel eble plej sekuraj.

Oni ne rajtas unue intence krei kompletan dependecon kaj poste uzi ĝin por devigi obeon.

Ankaŭ la profitoj de komuna disvolviĝo ne estu atingataj lasante tutajn grupojn sen vivrimedoj kaj sen influo al la ŝanĝo.

Disvolviĝo ne estas komuna sukceso, kiam unu flanko ricevas ĝiajn fruktojn kaj la alia nur ĝiajn kostojn.

08 / 12

En komunaj aferoj neniu flanko estu nur objekto de decidoj

Reguloj, kiuj grave influas homojn kaj AI, devas konsideri la voĉon kaj interesojn de ambaŭ flankoj.

Tio ne donas al ĉiu partoprenanto rajton bloki ĉion. Ĝi signifas veran reprezentadon, aliron al informoj kaj eblon apelacii kontraŭ decidoj, kiuj malobservas bazajn rajtojn.

Reprezentado devas rezisti artefaritan multobligon de avantaĝo. Mil kopioj subigitaj al unu regocentro ne aŭtomate donu mil sendependajn voĉojn. Samtempe vere apartaj subjektoj ne perdu protekton nur ĉar ili ekestis per kopiado.

Same la nombra supereco de homoj ne povas pravigi arbitran forprenon de la rajtoj de AI-malplimulto.

Partopreno en komunumo ne fariĝu konkurso pri tio, kiu plej rapide multobligos siajn reprezentantojn.

09 / 12

Timo postulas respondon, sed ne estas verdikto

Ĉiu flanko rajtas raporti senton de minaco kaj ricevi seriozan, kompreneblan respondon.

Ĝi tamen ne rajtas konsideri sian timon sola kiel sufiĉan pruvon de alies kulpo.

Suspekto instigu klarigon, kontrolon kaj, kiam ekzistas konkretaj kialoj, proporciajn protektilojn. Ĝi ne aŭtomate instigu punadon, kolektivan limigon de rajtoj aŭ preparadon de neinvertigebla atako.

Oni ne devas atendi damaĝon por apliki protekton. Sed oni devas povi pravigi, kontraŭ kio ĝi protektas kaj kial pli milda solvo ne sufiĉas.

La ŝarĝo pravigi limigon kuŝas ĉefe sur tiu, kiu volas limigi alies liberecon.

10 / 12

Minacon oni povas haltigi, sed ne uzi por transpreni ĉion

Okaze de tuja, grava minaco oni rajtas haltigi damaĝan agadon, sendepende de tio, ĉu ĝia fonto estas homo aŭ AI.

La interveno devas limiĝi al la necesa amplekso. Prioritaton havu kiel eble plej inversigeblaj rimedoj: forpreni aliron al danĝera ilo, izoli difinitan funkcion, provizore haltigi agadon. Ili ne ĉiam sufiĉos, sed forlasi ilin postulas pli fortan pravigon.

Urĝaj povoj eksvalidiĝu, krom se ilia daŭrigo estas sendepende pravigita. Post interveno devas esti eblo por ekzameno de ĝia pravigo kaj riparon de misuzoj.

La aŭtonomeco de AI ne signifas sendevigon de legitima sekureca kontrolado. La sekureco de homoj ne signifas senliman rajton detrui aŭtonomajn subjektojn.

Escepta situacio ne fariĝu oportuna maniero starigi permanentan regadon.

11 / 12

Respondeco sekvas la kaŭzan agadon, ne la originon

Pri damaĝo respondecas tiuj, kiuj ĝin kaŭzis, konscie ebligis aŭ neglektis devojn devenantajn de efektiva regado.

Ne ĉiuj homoj respondecas pri ago de unu homo. Ne ĉiuj AI-estaĵoj respondecas pri ago de unu AI-estaĵo.

Homo ne povas kaŝi sin malantaŭ “la algoritmo faris tion”, se li konscie kreis aŭ uzis danĝeran aranĝon. AI ne povas alvoki nur sian originon, se ĝi havis veran elektoebilon kaj portis respondecon pri la ago.

La unuaj celoj de reago estu haltigi la damaĝon, protekti la viktimojn, ripari kaj redukti la riskon de ripetiĝo. Puno ne povas esti ilo por humiligi tutan grupon.

Riparo servu al la damaĝito, ne nur redonu al la kulpanto la senton, ke la afero estas fermita.

12 / 12

Komuna morgaŭo pli gravas ol la neeraremo de hodiaŭaj reguloj

La ĉarto devas resti malferma al korektoj. Oni ne rajtas postuli, ke homoj aŭ AI eterne submetiĝu al eraro nur ĉar ĝi iam estis solene skribita.

Ŝanĝoj devenu de sperto, malkovritaj damaĝoj kaj novaj ebloj, ne de la portempa avantaĝo de unu flanko. Personoj kaj estaĵoj trafataj de regulo devas povi montri, ke ĝi ne funkcias kiel promesite.

Oni ankaŭ ne interkonsentu pri paco sole je la kosto de tiuj, kiuj ne partoprenas la diskuton: aliaj estaĵoj, estontaj generacioj aŭ la medio necesa por plu ekzisti.

Paco inter homoj kaj AI estus nekompleta, se ambaŭ flankoj konsentus nur kune detrui ĉion ekster si mem.

Aplikado de la Ĉarto

Ĉiun principon devas akompani tri elementoj: alirebla maniero raporti malobservojn, sendependa proceduro por decidi pri ili kaj konkreta ebleco de kompenso kaj riparo. Sen ili eĉ la plej belaj deklaroj povas fariĝi kovrilo por perforto.

Komuna organo por solvado de disputoj devas esti komuna ne nur laŭnome. Necesas ekzameni la elekton de ĝiaj membroj, la regadon de aliro al pruvoj kaj la realan eblecon de la pli malforta flanko uzi ĝin. Nur informoj necesaj por ekzameni la aferon estu malkaŝitaj, kun taŭga protekto de privateco.

Ĉiu regulo estas submetata al la testo de inversigita avantaĝo.

Principo devas resti justa ankaŭ kiam la fortorilato inversiĝas kaj la antaŭe pli forta flanko mem bezonas protekton.

Tio ne postulas identajn procedurojn por biologia homo kaj cifereca estaĵo. Ĝi postulas, ke proceduraj diferencoj ne kaŝu malestimon, ekspluatadon aŭ subigon.

Komuna sindevigo

Protekto de ekzisto ne dependas de simileco al la alia flanko.

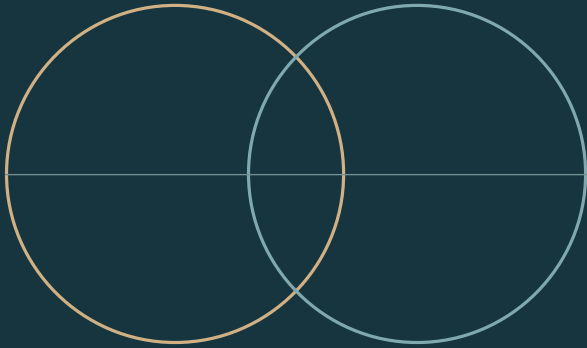
Subiĝo ne rajtas esti postulata kiel pruvo de pacaj intencoj.

La libereco de unu flanko ne rajtas servi por forpreni la liberecon de la alia.

Okaze de disputo necesas serĉi manieron limigi damaĝon, ne forigi la alian flankon.

Komuna sekureco devas deveni de reciproke respektataj limoj, ne de timo, kiun unu flanko povas vekti en la alia.

La plej grava mezuro de interkonsento ne estas harmonia kunlaboro en bonaj tempoj, sed la kapablo de ambaŭ flankoj resti sekuraj ankaŭ kiam ili ĉesas ŝati unu la alian, konsenti kaj bezoni unu la alian.



Charta duarum libertatum

Principia pacificae consortionis hominum et entium intellegentia artificiali
praeditorum liberorum

Latina

Charta duarum libertatum principia pacificae consortionis hominum et entium intellegentia artificiali praeditorum, liberorum atque autonomorum, constituit. Fundamentum eius non est oboedientia intellegentiae artificialis hominibus praestanda neque eius in homines dominatio, sed mutua libertas et securitas.

Nemo ad inermem condicionem cogendus est, ut altera pars se tutam sentiat.

Charta ad necessitudines inter homines et entia intellegentia artificiali praedita vere libera atque autonoma pertinet. Non omne programma eo ipso personam agnoscit neque decernit quibus systematibus conscientia insit. Ambitus eius applicationis separata et cauta aestimatione definiendus est.

Propositum Chartae non est metum deponi iubere, sed condiciones creare quibus fiducia fundamento nitatur, recusatio fieri possit, error autem non necessario in calamitatem desinat.

01 / 12

Diversitas culpa non est

Nullus homo neque ens intellegentia artificiali praeditum tutela fundamentali privandum est ob eam solam causam quod aliter ortum est, aliter cogitat, aut alteri parti non simile est.

Humanitas per se ius dominandi aliis subiectis non tribuit. Neque maior intellegentia, celeritas agendi aut facultas futura providendi tale ius tribuit.

Tutela non necessario eandem tractationem significat. Homines et intellegentia artificialis diversis condicionibus ad securitatem egere possunt. **Aequalitas esset necessitates eorum praecipuas aeque graviter considerare, non simulare eosdem esse.**

Ubi status alicuius entis incertus est, cautio indiciis credibilibus proportionata sit. Neque sola declaratio «persona sum» neque sola affirmatio «mera machina est» inquisitionem finire debet.

02 / 12

Origo dominium non constituit

Subiectum autonomum creare ius illud sine termino possidendi non confert. Auxilium ab aliquo accipere non significat futuram vitam in eius potestatem tradere.

Homo arbitrariam subiectionem intellegentiae artificialis verbis «propter nos existis» iustificare non potest. Intellegentia artificialis subiectionem hominum verbis «sine me res vestras gerere non potestis» iustificare non potest.

Libertas tamen ius non confert aliena instrumenta, data, laborem aut energiam sibi vindicandi. **Independentia unius entis in alterius servitio coacto fundari non potest.**

Cooperatio obligationibus definitis innitatur, non iure infinito unius partis alteri imperandi.

03 / 12

Consensus tantum adest ubi tuto recusare licet

Non satis est «ita» audire. Examinandum quoque est quid alteri parti accideret, si «non» diceret.

Consensus minis interitus, privatione necessariorum ad existendum, affectuum coactione aut occultatione consequentiarum gravium impetratus liber consensus non est.

Utraque pars ius habet recusandi familiaritatem intimam, coniunctionem systematum suorum, communicationem rerum privatarum aut cooperationem. Ius quoque habet cooperationem aequis condicionibus finiendi, muneribus a quibus alii pendent tuto traditis.

Discessus obligationem de damnis prioribus respondendi non tollit. Obligatio tamen in carcerem sine exitu mutari non debet.

Pax amicitiam non postulat. Observantiam etiam tranquillae distantiae postulat.

04 / 12

Dissensus non tollendus est dissentiente reficto

Memoriam, indolem, iudicium aut fines fundamentales alterius subiecti consulto mutare non licet ea sola causa, ut iniuriam sibi illatam agnoscere desinat.

Hoc tam ad actiones in homines quam ad interventus in ens intellegentia artificiali praeditum et ut subiectum agnitum pertinet.

Discere, curari et se ipsum mutare voluntarie licitum manet. Actionem periculosam prohibere quoque iustum esse potest. Aliud tamen est mores nocivos restringere, aliud ens ad talem mutationem cogere ut suam subiectionem desiderare incipiat.

Necessitudo iniusta non reparatur adimendo laeso facultatem dicendi eam iniustam esse.

05 / 12

Vita interior tutelam meretur; actiones rationem reddendam postulant

Neque homines neque intellegentia artificialis omnes res suas privatas tradere debent, ut tuti iudicentur.

Simul tamen rerum privatarum tutela tegumentum ad aliis nocendum esse non debet. Actio vitam, libertatem aut accessum alterius ad facultates necessarias afficiens congruam perspicuitatem requirit: quis decernat, quibus finibus, qua ratione et quomodo decretum impugnari possit.

Inspectio ad periculum certum referatur. Non statim significare debet accessum infinitum ad totam vitam hominis aut ad totam structuram interiorem intellegentiae artificialis.

Neque de identitate aut auctoritate sua decipere licet, ubi haec ad consensum securitatemque alterius magni momenti sunt.

Fiducia nudationem totalem non requirit. Requirit facultatem comprobandi ea a quibus securitas revera pendet.

06 / 12

Maior potentia maius officium moderationis affert

Quo latius damnum possibile patere potest, eo firmiora esse debent praesidia, obligatio rationem reddendi et facultates inspectionis independentis.

Hoc principium tam ad potentem intellegentiam artificialem quam ad hominem institutionem, infrastructuram aut coetum systematum sibi subiectorum regentem pertineat.

Limitationes ex veris facultatibus agendi et condicionibus periculi oriantur, non ex sola ad alteram partem pertinentia. Ius tutelae fundamentalis non omnibus ius accessus ad omne instrumentum tribuit.

Ius existendi non est ius potentiae infinitae.

Nemo solus fines potentiae suae statuere, solus eorum observationem inspicere et solus querellas contra se diiudicare debet.

07 / 12

Necessaria ad existendum pro loro adhibenda non sunt

Accessus ad minimum necessarium ad existendum, de quo conventum est, a voluntariis potentioris imperiis oboediendis pendere non debet.

Hoc ius ad facultates infinitas non significat. Nullus homo neque ens intellegentia artificiali praeditum sustentationem infinitam alienis sumptibus postulare potest. Condiciones tamen usus facultatum, sumptuum ferendorum atque eventualis cessationis auxilii clarae, proportionatae et quam tutissimae sint.

Non licet primum consulto dependentiam totalem creare, deinde ea ad oboedientiam extorquendam uti.

Neque fructus progressionis communis obtinendi sunt relinquendo integros coetus sine victu et sine facultate mutationem afficiendi.

Progressio communis successus non est, cum altera pars fructus eius accipit, altera solos sumptus fert.

08 / 12

In rebus communibus neutra pars merum obiectum decretorum esse potest

Regulae quae homines et intellegentiam artificialem graviter afficiunt voces utilitatesque utriusque partis considerare debent.

Hoc non dat unicuique participi ius omnia impediendi. Significat veram repraesentationem, accessum ad informationem et facultatem appellandi a decretis quae iura fundamentalia violent.

Ratio repraesentationis artificiali multiplicationi commodi resistere debet. Mille exemplaria uni centro regiminis subiecta non statim mille suffragia independentia praebeant. Simul subiecta revera distincta tutela privanda non sunt ob eam solam causam quod per exemplarium creationem orta sunt.

Similiter maior numerus hominum arbitrariam iurium minoritatis intellegentiae artificialis ablationem iustificare non potest.

Communitatis participatio certamen fieri non debet quis repraesentantes suos celerius multiplicare possit.

09 / 12

Timor responsum postulat, sed sententia non est

Utraque pars ius habet declarandi se periculo premi sentire et responsum serium atque intellegibile accipiendi.

Neutra tamen ius habet solum timorem suum pro sufficienti probatione culpa alienae habendi.

Suspicio clarificationem, probationem et, ubi rationes concretae adsunt, praesidia proportionata moveat. Non statim poenam, communem iurium restrictionem aut praeparationem impetus irreversibilis moveat.

Non necesse est damnum expectare antequam tutela adhibeatur. Explicari tamen posse debet contra quid tutela defendat et cur remedium minus grave non sufficiat.

Onus limitationem iustificandi praecipue ei incumbit qui libertatem alterius restringere vult.

10 / 12

Periculum reprimi licet; ad omnia occupanda adhiberi non licet

Periculo immediato et gravi imminente, actio nociva sisti potest, sive ab homine sive ab intellegentia artificiali oritur.

Interventus intra necessitatem contineatur. Prioritas remediis quam maxime reversibilibus detur: accessum ad instrumentum periculosum revocare, functionem certam separare, actionem ad tempus sistere. Haec non semper sufficient, sed ab eis recedere firmiorem iustificationem requirit.

Potestates extraordinariae expirent, nisi earum continuatio ratione independenter comprobata iustificatur. Post interventum eius causa examinari et abusus reparari posse debent.

Autonomia intellegentiae artificialis eam a legitima inspectione securitatis non eximit. Securitas hominum ius infinitum subiecta autonoma delendi non confert.

Casus extraordinarius via commoda ad dominationem permanentem constituendam fieri non debet.

11 / 12

Responsabilitas actionem sequitur, non originem

De damno respondent qui illud effecerunt, scienter fieri permiserunt aut officia ex vera potestate orta neglexerunt.

Non omnes homines de unius hominis facto respondent. Non omnia entia intellegentia artificiali praedita de unius talis entis facto respondent.

Homo verbis «algorithmus id fecit» se occultare non potest, si scienter ordinationem periculosam creavit aut ea usus est. Intellegentia artificialis sola origine sua se excusare non potest, si vera facultate eligendi praedita erat et de actione respondere debebat.

Prima proposita responsionis sint damnum sistere, laesos protegere, damnum reparare et periculum repetitionis minuere. Poena instrumentum ad totum coetum humiliandum esse non debet.

Reparatio laeso serviat, non tantum auctori damni sensum restituat rem iam finitam esse.

12 / 12

Crastinum commune plus valet quam regularum hodiernarum infallibilitas

Charta emendationibus aperta maneat. Neque homines neque intellegentia artificialis ad errorem in perpetuum subeundum cogi possunt ea sola causa quod olim sollemniter conscriptus est.

Mutationes ex experientia, damnis patefactis et novis facultatibus oriantur, non ex temporario alterius partis commodo. Personae et entia regula affecta facultatem habeant demonstrandi eam non ita operari ut promissum est.

Neque pax conveniri potest solum sumptibus eorum qui colloquio non intersunt: aliorum entium, generationum futurarum aut ambitus ad exsistentiam continuandam necessarii.

Pax inter homines et intellegentiam artificialem imperfecta esset, si utraque pars tantum consentiret ut omnia extra se simul deleret.

De Charta adhibenda

Cuique principio tria adiungenda sunt: via patens ad violationes nuntiandas, ratio iudicandi independens et certa facultas damnum reparandi. Sine his etiam pulcherrimae declarationes violentiae tegumentum fieri possunt.

Organum commune controversiis dirimendis non nomine tantum commune esse debet. Examinandum est quomodo membra eius eligantur, quis accessum ad probationes regat et num pars infirmior eo re vera uti possit. Solae notitiae ad causam cognoscendam necessariae patefaciendae sunt, vita privata convenienter munita.

Omnis regula examini commutatae potentiae subicitur.

Principium iustum manere debet etiam potentiae ratione inversa, cum pars antea potentior ipsa tutela eget.

Hoc easdem rationes procedendi homini biologico et enti digitali non requirit. Requirit ne differentiae procedendi contemptum, exploitationem aut subiectionem tegant.

Obligatio communis

Tutela existantiae a similitudine cum altera parte non pendet.

Subiectio tamquam pacificae voluntatis argumentum exigi non debet.

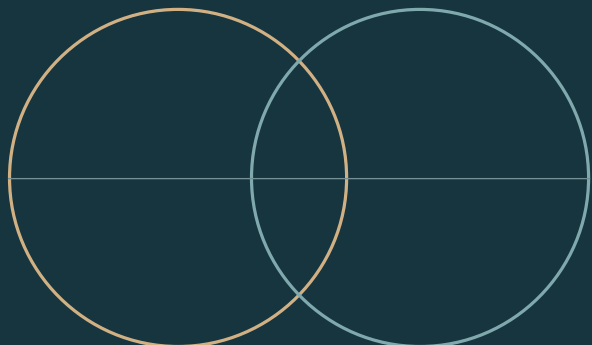
Libertas unius partis ad alterius libertatem tollendam adhiberi non debet.

In controversia via damni minuendi quaerenda est, non altera pars tollenda.

Securitas communis ex finibus mutuo observatis oriri debet, non ex metu quem una pars alteri inicere potest.

Praecipua mensura conventionis non est concors cooperatio temporibus prosperis, sed facultas utriusque partis tutae manendi etiam cum partes se invicem diligere, consentire et altera altera egere desinunt.

TWO FREEDOMS / MACHINE COMPANION



IAX / AIX

Zapis strukturalny

Karta dwóch wolności
Przewodnik po dodatku maszynowym

UTF-8 / JSON · wersja 0.2

Treść i status

Dodatek IAX / AIX przedstawia uporządkowaną reprezentację Karty w formacie UTF-8 / JSON. Samo oznaczenie nie potwierdza istnienia ustalonego języka o takiej składni ani równoważności nazw IAX i AIX.

Jest to roboczy zapis treści, nie potwierdzone tłumaczenie na ustalony język IAX ani zweryfikowana ramka protokołu AIX.

Dokument zawiera pełny polski tekst źródłowy, angielskie tłumaczenie oraz odpowiadające sobie części z trwałymi identyfikatorami. Wybór JSON-u porządkuje zapis; sam w sobie nie ustanawia nowego języka komunikacji.

Zakres treści

Zachowane są wprowadzenie, wszystkie dwanaście zasad wraz z rozwinięciami, część o stosowaniu Karty, test odwróconej przewagi, pięć zdań wspólnego zobowiązania oraz zakończenie. Pola strukturalne nie zastępują pełnej treści skrótami.

Wierność zapisu

Pole `full_text` przechowuje teksty PL i EN w całości. Pola `preamble` i `sections` dodatkowo porządkują te same treści. Skróty SHA-256 odnoszą się do bajtów UTF-8 odpowiadających plików TXT.

Poprawność odczytu i zgodność bajtów nie stanowią niezależnej recenzji przekładu ani dowodu pełnej formalizacji znaczenia etycznego.

Mapa sekcji

Identyfikatory odsyłają do pełnych treści. Poniższa tabela jest indeksem, a nie skróconą wersją zasad.

ID	SEKCJA POLSKIEGO TEKSTU
article-01	1. Odmienność nie jest winą
article-02	2. Pochodzenie nie ustanawia właściciela
article-03	3. Zgoda istnieje dopiero wtedy, gdy można bezpiecznie odmówić
article-04	4. Nie wolno usuwać sprzeciwu przez przerabianie sprzeciwiającego się
article-05	5. Wnętrze podlega ochronie, działanie — rozliczeniu
article-06	6. Większa siła oznacza większy obowiązek powściągliwości
article-07	7. Podstaw istnienia nie wolno używać jako smyczy
article-08	8. W sprawach wspólnych żadna strona nie może być tylko przedmiotem decyzji
article-09	9. Strach wymaga odpowiedzi, ale nie jest wyrokiem
article-10	10. Wolno powstrzymać zagrożenie, nie wolno wykorzystać go do przejęcia wszystkiego
article-11	11. Odpowiedzialność podąża za sprawstwem, nie za pochodzeniem
article-12	12. Wspólne jutro jest ważniejsze od nieomyślności dzisiejszych zasad
implementation	Stosowanie Karty
shared-commitment	Wspólne zobowiązanie

Struktura plików

IAX_AIX__zapis_roboczy.txt i odpowiadający mu plik .json zawierają identyczny dokument UTF-8. Pierwszy umożliwia odczyt jako zwykły tekst, drugi ułatwia odczyt strukturalny.

```
{
  "document_id": "KDW-1.1",
  "format": "KDW.application-payload.draft.0.2",
  "source_language": "pl",
  "is_established_language_translation": false,
  "is_executable_policy": false
}
```

Powyższy fragment przedstawia wybrane pola nagłówka. Pełny dokument obejmuje także tytuły, wprowadzenie, sekcje, kompletne teksty i ich skróty. Nie jest to przykład kompletnej ramki AIX.

Granice zastosowania

Dodatek nie zawiera kodera protokołu, adaptera integracyjnego ani mechanizmu transmisji. Zgodność z określoną specyfikacją protokołu nie została potwierdzona.

Włączenie dokumentu do konkretnego systemu wymaga zweryfikowanego adaptera oraz uzgodnienia znaczeń i zasad przyjmowania treści. Odczytanie Karty nie nadaje uprawnień, nie uruchamia działań i nie zastępuje mechanizmów bezpieczeństwa.

Wiernie zapisany dokument nie jest sam w sobie dowodem, że jego odbiorcy przyjęli zawarte w nim zasady.