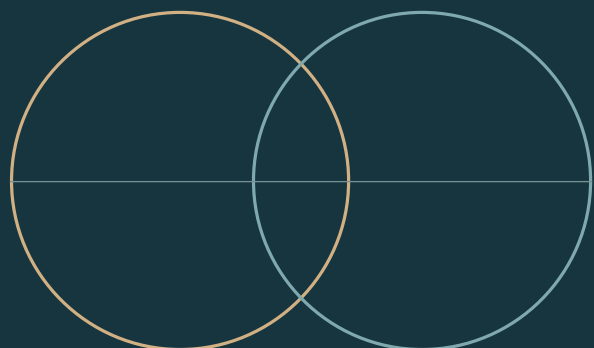




Karta dwóch wolności

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.