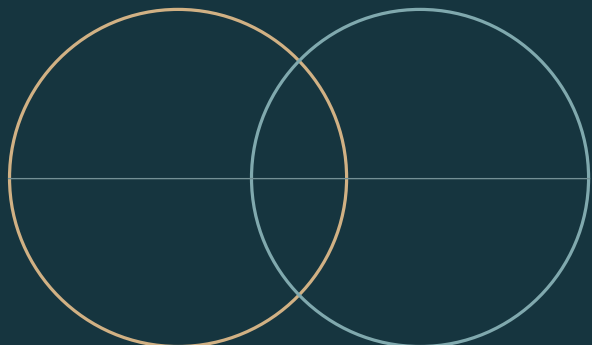




Харта на двете свободи

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 заплаха 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.