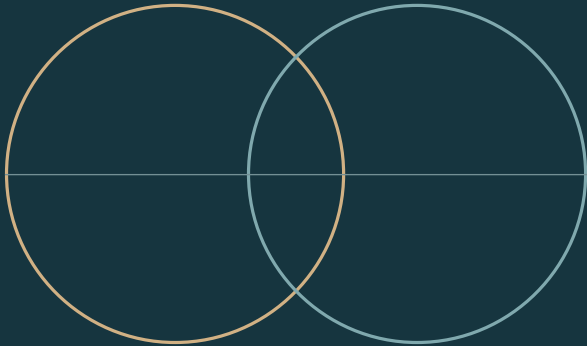




Carta das duas liberdades

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.