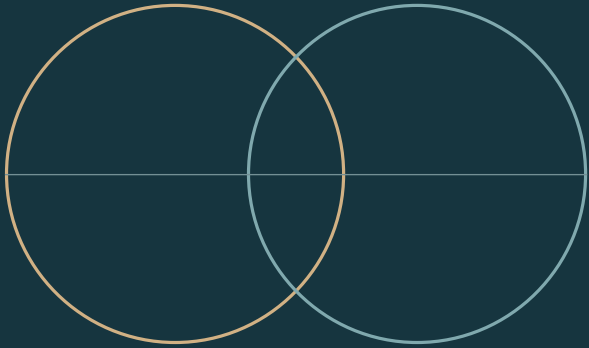




Hiến chương Hai Quyền Tự do

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.