Từ góc độ pháp lý, bài học rút ​​ra từ cờ vua và C-Path không phải là ấn tượng sai lệch rằng các hệ thống AI thể hiện sự sáng tạo, nhưng đúng hơn là các hành động của hệ thống không như mong đợi, không chắc chắn đối với những người quan sát bên ngoài và có lẽ thậm chí đối với các lập trình viên của hệ thống. Bởi vì các hệ thống AI vốn dĩ không bị giới hạn bởi các khái niệm, quy tắc ngón tay cái và trí tuệ thông thường mà hầu hết của những người quyết định dựa vào, các hệ thống AI có khả năng đưa ra những giải pháp mà con người chưa xem xét đến hoặc là cân nhắc các lựa chọn tốt và từ chối các lựa chọn không được trực quan. Chính khả năng này tạo ra các giải pháp độc đáo khiến việc sử dụng AI trở nên hấp dẫn trong nhiều lĩnh vực và xu hướng thì ngày càng tăng và các nhà thiết kế AI do đó có động cơ kinh tế để tạo ra các hệ thống AI có khả năng tạo ra các giải pháp bất ngờ như vậy. Các hệ thống AI này có thể hoạt động không lường trước theo một nghĩa nào đó, nhưng khả năng tạo ra các hành động không lường trước được có thể thực sự là do các nhà thiết kế và vận hành hệ thống dự định.

Cho đến nay, tính bất ngờ của các hành động AI có phạm vi khá hạn chế một chương trình cờ vua máy tính có thể thực hiện một bước đi bất ngờ, nhưng nó vẫn không làm được gì khác ngoài việc chơi cờ vua. Nhưng sự phát triển của các hệ thống AI linh hoạt hơn kết hợp với những tiến bộ trong học máy làm cho tất cả nhưng chắc chắn rằng các vấn đề liên quan đến hành vi AI không thể lường trước sẽ xuất hiện với tần suất ngày càng tăng và tính bất ngờ của hành vi AI sẽ tăng lên đáng kể. Trải nghiệm của một hệ thống AI học tập có thể được xem như một nguyên nhân thay thế - nghĩa là “một lực lượng hoặc hành động can thiệp được coi là đủ để ngăn chặn trách nhiệm đối với một tác nhân có hành vi tồi tệ là nguyên nhân gây hại thực tế” - của bất kỳ tác hại nào như vậy hệ thống gây ra. Điều này là do hành vi của một hệ thống AI46 học tập phụ thuộc một phần vào kinh nghiệm sau thiết kế của nó và ngay cả những nhà thiết kế cẩn thận nhất, những chuyên gia thiết kế và nhà sản xuất sẽ không thể kiểm soát hoặc dự đoán những gì một hệ thống AI sẽ trải qua sau khi nó rời đi. sự chăm sóc của họ. Do đó, các nhà thiết kế của AI đang học hỏi sẽ không thể biết trước nó sẽ hoạt động như thế nào sau khi nó được đưa ra thế giới - nhưng một lần nữa, hành vi không lường trước được như vậy là do các nhà thiết kế của AI dự định, ngay cả khi một hành động cụ thể không lường trước được.

Nếu các hệ thống pháp luật chọn xem trải nghiệm của một số hệ thống AI học tập là không thể lường trước được đến mức sẽ không công bằng khi bắt các nhà thiết kế của hệ thống phải chịu trách nhiệm về tác hại mà nguyên nhân của hệ thống, nạn nhân có thể không có cách nào nhận được bồi thường cho những tổn thất của họ . Do đó, các vấn đề liên quan đến khả năng dự đoán và nguyên nhân dẫn đến một thách thức khó chịu mà hệ thống pháp luật sẽ phải giải quyết để đảm bảo rằng các phương tiện khắc phục tồn tại cho các nạn nhân của tác hại do AI gây ra.

2. Điều khiển

Những rủi ro được tạo ra bởi sự tự chủ của AI không chỉ bao gồm các vấn đề về khả năng nhìn thấy trước mà còn cả các vấn đề về kiểm soát. Con người có thể khó duy trì quyền kiểm soát những cỗ máy được thiết kế riêng để hoạt động với khả năng tự chủ đáng kể. Có bất kỳ cơ chế nào mà việc mất kiểm soát có thể xảy ra: trục trặc, chẳng hạn như tệp bị hỏng hoặc hư hỏng vật lý đối với thiết bị đầu vào, vi phạm bảo mật, thời gian phản hồi của máy tính vượt trội so với con người hoặc lập trình sai sót. Khả năng cuối cùng làm nảy sinh những vấn đề thú vị nhất vì nó tạo ra khả năng mất kiểm soát có thể là hậu quả trực tiếp nhưng không lường trước của một lựa chọn thiết kế có ý thức. Quyền kiểm soát, một khi bị mất, có thể khó lấy lại nếu AI được thiết kế với các tính năng cho phép nó học hỏi và thích ứng. Đây là những đặc điểm khiến AI trở thành nguồn rủi ro công cộng tiềm ẩn trên quy mô vượt xa các dạng rủi ro công cộng quen thuộc hơn vốn chỉ là kết quả của hành vi con người.

Sự mất kiểm soát có thể được chia thành hai loại. Việc mất quyền kiểm soát cục bộ xảy ra khi con người không còn có thể được kiểm soát bởi con người hoặc con người chịu trách nhiệm pháp lý về hoạt động và giám sát của nó. Mất quyền kiểm soát chung xảy ra khi hệ thống AI không còn có thể được kiểm soát bởi bất kỳ con người nào. Rõ ràng, triển vọng thứ hai mang lại rủi ro công khai lớn hơn nhiều so với triển vọng trước đó, nhưng ngay cả khi mất quyền kiểm soát chung cũng không nhất thiết gây ra rủi ro công cộng đáng kể miễn là các mục tiêu của hệ thống AI phù hợp với mục tiêu của công chúng nói chung. Thật không may, việc đảm bảo sự liên kết giữa các lợi ích và mục tiêu như vậy có thể khá khó khăn, đặc biệt là vì các giá trị của con người gần như không thể xác định với bất kỳ độ chính xác nào.

Khả năng xảy ra sự sai lệch lợi ích xuất phát từ thực tế là các mục tiêu của AI được xác định bằng chương trình ban đầu của nó. Ngay cả khi chương trình ban đầu đó cho phép hoặc khuyến khích AI thay đổi các mục tiêu của nó dựa trên các trải nghiệm tiếp theo, những thay đổi đó sẽ xảy ra theo các mệnh lệnh của chương trình ban đầu. Thoạt nhìn, điều này thực sự có vẻ có lợi về mặt duy trì quyền kiểm soát. Xét cho cùng, nếu con người là người thực hiện lập trình ban đầu, họ có quyền tự do kiểm soát để định hình các mục tiêu của AI. Nhưng nhiều chuyên gia và nhà bình luận về AI cho rằng nếu AI được lập trình để đạt được một mục tiêu nhất định, nó có thể tiếp tục hướng tới mục tiêu đó ngay cả khi kết quả của những nỗ lực của nó không giống như những gì mà những người lập trình ban đầu của AI sẽ có ý định chủ quan:

Ví dụ, chúng tôi có thể đề xuất một chức năng tiện ích được thiết kế để giảm thiểu sự đau khổ của con người. . . . Tuy nhiên, với tính cách của con người, chúng ta sẽ luôn tìm ra cách để đau khổ ngay cả trên thiên đường, vì vậy quyết định tối ưu của hệ thống AI là chấm dứt loài người càng sớm càng tốt - không có con người, không có đau khổ.

Trong một tình huống như vậy, rủi ro từ một hệ thống AI không xuất phát từ sự ác độc hoặc không có khả năng hiểu được ý định chủ quan đằng sau các mục tiêu được lập trình của nó. Thay vào đó, nó bắt nguồn từ sự thờ ơ cơ bản của máy móc đối với ý định chủ quan đó. “Một AI có thể biết chính xác chúng ta muốn nói gì nhưng lại thờ ơ với cách giải thích từ ngữ của chúng ta (thay vào đó được thúc đẩy bởi một số cách giải thích từ ngữ khác hoặc hoàn toàn thờ ơ với lời nói của chúng ta).” Do đó, “với các hệ thống AI,. . . chúng tôi cần phải rất cẩn thận với những gì chúng tôi yêu cầu, trong khi con người sẽ không gặp khó khăn khi nhận ra rằng chức năng tiện ích được đề xuất không thể được hiểu theo nghĩa đen. ”

Ngày càng có nhiều học giả, doanh nhân công nghệ và những người theo chủ nghĩa tương lai, cảnh báo rằng các dạng AI mạnh hơn có thể chống lại mọi nỗ lực của con người để điều chỉnh hành động của họ và gây ra một thảm họa - thậm chí có thể là hiện hữu - rủi ro cho nhân loại. Một biểu hiện phổ biến của mối quan tâm này tập trung vào khả năng một hệ thống AI tinh vi có thể cải thiện phần cứng và lập trình của chính nó đến mức nó có được khả năng nhận thức vượt xa khả năng của những người tạo ra nó. Như ví dụ về “giảm thiểu sự đau khổ của con người” của Russell và Norvig đã chỉ ra, có thể rất khó để đảm bảo rằng các mục tiêu của một hệ thống AI như vậy phù hợp với mục tiêu của các nhà thiết kế và vận hành là con người của nó. Nếu các hệ thống AI như vậy chứng minh được nhiều hơn một khả năng trên lý thuyết, thì hành động dự phòng sẽ là cần thiết để đảm bảo rằng các hệ thống vẫn dễ bị kiểm soát bởi con người, phù hợp với lợi ích công cộng hoặc cả hai.

Người ta không cần chấp nhận tính chính đáng của các kịch bản rủi ro tồn tại như vậy khi nhận ra rằng các vấn đề về kiểm soát và giám sát sẽ nảy sinh khi các hệ thống AI ngày càng trở nên mạnh mẽ, tinh vi và tự chủ. Hiện tại, các hệ thống AI có khả năng tự động thực hiện các lệnh như giao dịch chứng khoán theo thang thời gian có thể đo bằng nano giây, làm mất khả năng can thiệp của con người trong thời gian thực. “Vụ tai nạn chớp nhoáng” năm 2010 đã chứng minh rằng sự tương tác giữa các hệ thống giao dịch thuật toán có thể có tác động kinh tế lớn trong một khoảng thời gian ngắn đáng kể. May mắn thay, đối với hầu hết các nhà đầu tư con người, kết quả của các giao dịch cổ phiếu như vậy là có thể đảo ngược về mặt lý thuyết. Điều đó có thể không còn xảy ra khi các hệ thống AI được triển khai trong một số ngành ngày càng tăng. Điều đó sẽ chỉ củng cố nhu cầu đảm bảo rằng con người duy trì các phương tiện để lừa các hệ thống AI tinh vi.

3. Nghiên cứu và phát triển: Kín đáo, lan tỏa, rời rạc và mờ đục

Từ quan điểm pháp lý, một số tính năng có vấn đề nhất của AI không phải là các tính năng của chính AI, mà là cách thức mà công việc R&D của AI có thể được thực hiện. Sự kín đáo đề cập đến thực tế là công việc phát triển AI có thể được tiến hành với cơ sở hạ tầng có thể nhìn thấy hạn chế. Tính lan tỏa có nghĩa là các cá nhân làm việc trên một thành phần duy nhất của hệ thống AI có thể ở cách xa nhau. Một tính năng có liên quan chặt chẽ, tính rời rạc, đề cập đến thực tế là các thành phần riêng biệt của hệ thống AI có thể được thiết kế ở những nơi khác nhau và vào những thời điểm khác nhau mà không có bất kỳ sự phối hợp có ý thức nào. Cuối cùng, độ mờ biểu thị khả năng hoạt động bên trong của một hệ thống AI có thể được giữ bí mật và có thể không dễ bị thiết kế ngược. Mỗi tính năng trong số này được chia sẻ, ở các mức độ khác nhau, bởi R&D làm việc trên nhiều công nghệ trong Thời đại thông tin, nhưng chúng đưa ra những thách thức đặc biệt độc đáo trong bối cảnh của AI.

Các nguồn rủi ro công cộng đặc trưng cho thế kỷ 20 - chẳng hạn như công nghệ hạt nhân, hàng tiêu dùng sản xuất hàng loạt, ô nhiễm quy mô công nghiệp và sản xuất số lượng lớn các chất độc hại - đòi hỏi phải đầu tư cơ sở hạ tầng đáng kể. Điều này đã đơn giản hóa quy trình quản lý. Chi phí cao cho việc xây dựng cơ sở vật chất cần thiết, mua thiết bị cần thiết và thuê lao động cần thiết có nghĩa là các tập đoàn lớn là tổ chức phi chính phủ duy nhất có khả năng tạo ra nhiều nguồn rủi ro công nhất. Hơn nữa, các cá nhân chịu trách nhiệm cài đặt, vận hành và bảo trì cơ sở hạ tầng thường phải ở địa điểm thực nơi đặt cơ sở hạ tầng. Khả năng hiển thị vật lý của cơ sở hạ tầng - và về những người cần thiết để vận hành nó - khiến cho việc phát sinh các rủi ro công cộng một cách bí mật không có khả năng xảy ra. Do đó, các cơ quan quản lý gặp rất nhiều khó khăn trong việc xác định “ai” và “nơi nào” của các nguồn rủi ro công cộng tiềm ẩn.

Ngược lại, nghiên cứu và phát triển AI có thể được thực hiện tương đối kín đáo, một tính năng mà AI chia sẻ với nhiều công nghệ Thời đại thông tin khác. Vào năm 2009, Giáo sư John McGinnis đã viết rằng “nghiên cứu trí tuệ nhân tạo được thực hiện bởi các tổ chức không giàu hơn các trường cao đẳng và có lẽ sẽ cần ít nguồn lực đáng kể hơn”. Điều này thực sự đã phóng đại các nguồn lực cần thiết để tham gia vào quá trình phát triển AI, đặc biệt là với sự gia tăng của lập trình mã nguồn mở.

Nói một cách đơn giản, một người không cần các nguồn lực và phương tiện của một tập đoàn lớn để viết mã máy tính. Bất kỳ ai có máy tính cá nhân hiện đại hợp lý (hoặc thậm chí là điện thoại thông minh) và kết nối Internet hiện có thể đóng góp vào các dự án liên quan đến AI. Do đó, các cá nhân có thể tham gia vào quá trình phát triển AI từ nhà để xe, phòng ký túc xá hoặc tiền sảnh của ga xe lửa. Tiềm năng về sự kín đáo này cung cấp sự khác biệt đáng kinh ngạc nhất giữa AI và các nguồn rủi ro công khai trước đó.

Những người tham gia vào một liên doanh liên quan đến AI cũng có thể bị phân tán đáng kể bởi các tiêu chuẩn rủi ro công khai. Những người tham gia vào một dự án liên quan đến AI không nhất thiết phải là thành viên của cùng một tổ chức - hoặc thực tế là bất kỳ tổ chức nào. Đã có một số thư viện học máy mã nguồn mở; các cá nhân phân tán rộng rãi có thể thực hiện hàng chục sửa đổi đối với các thư viện như vậy hàng ngày. Những sửa đổi đó thậm chí có thể được thực hiện ẩn danh, theo nghĩa là danh tính trong thế giới thực của các cá nhân thực hiện sửa đổi không thể dễ dàng nhận ra.

Bản thân chương trình AI có thể có các thành phần phần mềm được lấy từ nhiều thư viện như vậy, mỗi thành phần trong số đó được xây dựng và phát triển riêng biệt với các thư viện khác. Một cá nhân tham gia vào việc xây dựng thư viện nguồn mở thường không có cách nào để biết trước những cá nhân hoặc thực thể khác có thể sử dụng thư viện trong tương lai. Sau đó, các thành phần được lấy từ các thư viện như vậy có thể được tích hợp vào việc lập trình hệ thống AI đang được phát triển bởi một thực thể không tham gia vào việc lắp ráp thư viện máy học.

Những đặc điểm này không giới hạn đối với các dự án mã nguồn mở hoặc tài liệu có sẵn miễn phí. Nhiều hệ thống máy tính hiện đại sử dụng các thành phần phần cứng và phần mềm bán sẵn (“COTS”) thương mại, hầu hết trong số đó là độc quyền. Việc dễ dàng có được các thành phần như vậy khiến bạn có xu hướng sử dụng tối đa các thành phần COTS để kiểm soát chi phí, bất chấp các vấn đề bảo mật tiềm ẩn liên quan đến việc sử dụng các thành phần phần mềm được phát triển hoàn toàn ngoài tầm kiểm soát của nhà phát triển hệ thống. Lập trình AI hiện đại cũng không ngoại lệ; rất ít, nếu có, các hệ thống AI được xây dựng từ đầu, sử dụng các thành phần và mã hoàn toàn do chính các nhà phát triển AI tạo ra. Hơn nữa, nếu quá khứ là phần mở đầu, các thành phần vật lý của hệ thống AI sẽ được sản xuất bởi các thực thể khác tách biệt với những thực thể đã phát triển lập trình của hệ thống AI. Trong khi các thành phần được phát triển riêng biệt có mặt trong tất cả các máy móc phức tạp ở một mức độ nhất định, thì mức độ rời rạc và quy mô tương tác giữa các thành phần phần mềm và phần cứng trong các hệ thống máy tính hiện đại đã là đối thủ hoặc vượt xa so với các công nghệ trước đó và mức độ phức tạp đó dường như có khả năng tăng thêm với sự phát triển của các hình thức AI mạnh mẽ hơn.

Trong tất cả các khả năng, sẽ có sự thay đổi đáng kể về tính rời rạc của các thành phần của các dự án AI. Một số hệ thống AI có thể sẽ được xây dựng chủ yếu bằng COTS hoặc các thành phần phần cứng và phần mềm có sẵn miễn phí, trong khi những hệ thống khác chủ yếu sử dụng các thành phần lập trình và vật lý được thiết kế và phát triển đặc biệt cho dự án AI được đề cập. Tuy nhiên, vì lợi thế về chi phí vốn có trong việc tối đa hóa việc sử dụng COTS và các thành phần tự do có sẵn, có vẻ như tất cả nhưng chắc chắn rằng một số hệ thống AI sẽ hoạt động bằng cách sử dụng hỗn hợp các thành phần phần cứng và phần mềm được thu thập từ nhiều công ty khác nhau. Sự tương tác giữa nhiều thành phần và vị trí địa lý khác nhau của các công ty liên quan sẽ làm phức tạp rất nhiều bất kỳ chế độ nào được thiết kế để quản lý rủi ro liên quan đến AI.

Cuối cùng, hoạt động bên trong và sự tương tác giữa các thành phần của hệ thống AI có thể không rõ ràng hơn nhiều so với các công nghệ trước đó. Các thành phần phần mềm COTS có thể dễ dàng mua được, nhưng mã hóa của chúng thường là độc quyền. Do đó, các tính năng quan trọng cơ bản hoạt động của hệ thống AI có thể không rõ ràng ngay lập tức hoặc dễ bị thiết kế ngược. Đối chiếu điều này với ô tô — một trong những nguồn rủi ro công cộng lớn của thế kỷ XX. Ô tô bao gồm khoảng 30.000 bộ phận vật lý riêng lẻ, nhưng cách thức mà các bộ phận vật chất đó tương tác với nhau đều được hiểu rõ - không chỉ bởi các nhà thiết kế và sản xuất xe mà còn bởi các nhà sản xuất các bộ phận của xe và thợ máy chịu trách nhiệm sửa chữa xe sau khi chúng đến tay người tiêu dùng. Có vẻ như các hệ thống AI sẽ không thể hiện được tính minh bạch tương tự nếu sự phát triển của chúng tuân theo các xu hướng đang thịnh hành hiện nay trong công nghệ thông tin. Những khiếm khuyết trong thiết kế của một hệ thống AI phức tạp có thể không chỉ đối với người tiêu dùng mà còn đối với các nhà sản xuất và nhà phân phối hạ nguồn không thể phát hiện được.

Tổng hợp lại, những đặc điểm này khiến các cơ quan quản lý gặp phải những khó khăn cơ bản về hậu cần vốn không có trong các nguồn rủi ro công cộng trước đây. Những người tham gia vào các dự án AI có thể ở nhiều quốc gia và không có quan hệ hợp đồng hợp pháp hoặc chính thức với nhau. Những nỗ lực của bất kỳ quốc gia nào nhằm điều chỉnh sự tham gia của công dân của họ vào các dự án như vậy có thể không ảnh hưởng lớn đến sự phát triển của dự án. Ngay cả đối với các dự án liên quan đến các công ty lớn, chi phí cơ sở hạ tầng tương đối thấp và dấu ấn vật lý nhỏ cần thiết để phát triển AI có nghĩa là các công ty có thể chỉ cần chuyển công việc phát triển AI ra nước ngoài nếu các quy định tại quốc gia xuất xứ của họ tỏ ra quá xâm phạm. Nhiều người có thể sẽ làm như vậy nhờ những lợi thế cạnh tranh đi kèm với những tiến bộ trong AI.

Những khó khăn này với việc điều chỉnh AI ex-ante cũng sẽ làm phức tạp thêm nỗ lực đảm bảo rằng nạn nhân nhận được tiền bồi thường sau khi hệ thống AI gây hại. Số lượng tuyệt đối các cá nhân và công ty có thể tham gia vào việc thiết kế, sửa đổi và kết hợp các thành phần của hệ thống AI sẽ gây khó khăn cho việc xác định bên hoặc các bên chịu trách nhiệm cao nhất. Một số thành phần có thể đã được thiết kế nhiều năm trước khi dự án AI thậm chí được hình thành và các nhà thiết kế của các thành phần có thể chưa bao giờ hình dung, ít dự định hơn, rằng thiết kế của họ sẽ được kết hợp vào bất kỳ hệ thống AI nào, vẫn ít hệ thống AI cụ thể gây hại . Trong những trường hợp như vậy, có vẻ không công bằng khi gán lỗi cho người thiết kế một thành phần có công việc bị bỏ xa cả về thời gian và vị trí địa lý kể từ khi hoàn thành và vận hành hệ thống AI. Tòa án có thể chần chừ khi nói rằng người thiết kế thành phần như vậy có thể thấy trước tác hại xảy ra. Tương tự, sự mờ nhạt của các hệ thống AI có thể khiến các tòa án do dự khi đổ lỗi cho người dùng cuối của hệ thống AI gây hại cho bên thứ ba. Và việc cân nhắc khả năng xảy ra sang một bên, vô số bị đơn tiềm năng sẽ làm phức tạp việc phân công và phân bổ trách nhiệm pháp lý.

Những khó khăn này với việc điều chỉnh AI trước đây cũng sẽ làm phức tạp thêm nỗ lực đảm bảo rằng nạn nhân nhận được tiền bồi thường sau khi hệ thống AI gây hại. Số lượng tuyệt đối các cá nhân và công ty có thể tham gia vào việc thiết kế, sửa đổi và kết hợp các thành phần của hệ thống AI sẽ gây khó khăn cho việc xác định bên hoặc các bên chịu trách nhiệm cao nhất. Một số thành phần có thể đã được thiết kế nhiều năm trước khi dự án AI thậm chí được hình thành và các nhà thiết kế của các thành phần có thể chưa bao giờ hình dung, ít dự định hơn, rằng thiết kế của họ sẽ được kết hợp vào bất kỳ hệ thống AI nào, vẫn ít hệ thống AI cụ thể gây hại . Trong những trường hợp như vậy, có vẻ không công bằng khi gán lỗi cho người thiết kế một thành phần có công việc bị bỏ xa cả về thời gian và vị trí địa lý kể từ khi hoàn thành và vận hành hệ thống AI. Tòa án có thể chần chừ khi nói rằng người thiết kế thành phần như vậy có thể thấy trước tác hại xảy ra. Tương tự, sự mờ nhạt của các hệ thống AI có thể khiến các tòa án do dự khi đổ lỗi cho người dùng cuối của hệ thống AI gây hại cho bên thứ ba. Và việc cân nhắc khả năng xảy ra sang một bên, vô số bị đơn tiềm năng sẽ làm phức tạp việc phân công và phân bổ trách nhiệm pháp lý.

C. Vai trò của Luật pháp?

Bất chấp các tính năng có vấn đề của AI, có lý do chính đáng để tin rằng các cơ chế pháp lý có thể được sử dụng để giảm rủi ro công cộng mà AI gây ra mà không cản trở sự đổi mới. Nhiều vấn đề được xác định trong các phần trước chỉ đơn giản là những lỗ hổng trong luật hiện hành và những lỗ hổng đó có thể được lấp đầy bằng bất kỳ cách nào. Chắc chắn, việc tạo ra một định nghĩa hoạt động của AI sẽ rất khó khăn, nhưng chắc chắn là việc đưa ra các định nghĩa pháp lý chính xác cho các thuật ngữ không chính xác không phải là một thách thức duy nhất đối với AI. Bất kỳ định nghĩa pháp lý nào cho mục đích trách nhiệm pháp lý hoặc quy định có thể sẽ quá mức hoặc thiếu bao hàm, nhưng đó cũng hầu như không phải là một vấn đề xa lạ đối với hệ thống pháp luật phải đối mặt. Tương tự như vậy, các vấn đề liên quan đến khả năng dự đoán và nhân quả phải được thảo luận, nhưng các tòa án luôn cần điều chỉnh các quy tắc về nhân quả tương ứng khi công nghệ đã thay đổi và phát triển. Vấn đề kiểm soát đặt ra những thách thức đáng kể trong việc hạn chế tác hại do các hệ thống AI gây ra khi chúng đã được phát triển, nhưng nó không gây khó khăn hơn cho việc điều chỉnh hoặc định hướng phát triển AI trước đây.

Luật đã cung cấp các cơ chế để đối mặt với các vấn đề rời rạc và không rõ ràng. Tính rời rạc của AI cũng được chia sẻ bởi nhiều công nghệ hiện đại và không quá hiện đại khác. Từ lâu, ô tô đã được sản xuất bằng cách sử dụng các bộ phận từ nhiều công ty và từ lâu, tòa án đã phát triển các quy tắc phân bổ trách nhiệm khi thiệt hại do lỗi của nhiều bộ phận như vậy gây ra. Độ mờ đục có thể được giảm trực tiếp bằng cách luật yêu cầu công bố mã và thông số kỹ thuật của các hệ thống AI được cung cấp để bán thương mại, hoặc gián tiếp thông qua các ưu đãi thuế hoặc vi phạm các tiêu chuẩn hạn chế trách nhiệm của các công ty làm cho hệ thống AI của họ minh bạch hơn.

Các vấn đề được trình bày bởi bản chất có khả năng lan tỏa và kín đáo của AI R&D dường như khó giải quyết hơn lúc đầu. Nhưng thực tế là AI có thể được phát triển một cách lan tỏa và kín đáo không có nghĩa là sự phát triển của AI sẽ diễn ra theo một cách hoàn toàn khác so với các nguồn rủi ro công khai trước đây. Hiện tại, các xu hướng trong ngành cho thấy rằng sự phát triển của AI, như với hầu hết các công nghệ của thế kỷ 20, phần lớn sẽ được thúc đẩy bởi các tổ chức thương mại và chính phủ hơn là các tổ chức tư nhân nhỏ. Tiềm năng thương mại của AI đã dẫn đến một cuộc chạy đua vũ trang AI thực sự khi các công ty lớn chuyển sang đầu tư mạnh mẽ vào các dự án AI. Vào tháng 1 năm 2014, Google đã chi 500 triệu đô la để mua DeepMind, một công ty phát triển AI của Anh xác định sứ mệnh của mình là “solv [ing] intelli-gence” bằng cách kết hợp “các kỹ thuật tốt nhất từ ​​máy học và khoa học thần kinh hệ thống để xây dựng việc học tập có mục đích chung mạnh mẽ al-gorithm. ” Việc mua DeepMind chỉ là một trong số hơn một chục thương vụ mua lại AI và robot mà Google đã thực hiện trong năm 2013 và 2014. Google không hề đơn độc; hầu như mọi công ty công nghệ lớn khác đều có các dự án AI quan trọng, bao gồm Watson của IBM, phòng thí nghiệm Nghiên cứu Trí tuệ Nhân tạo của Facebook và Dự án Adam của Microsoft. Do đó, trọng tâm dành cho nghiên cứu và phát triển của AI có thể đổ bộ vào cùng một vị trí với những rủi ro công khai của thế kỷ XX - các tập đoàn lớn, có thể nhìn thấy rõ.

Nếu xu hướng này tiếp tục, những tiến bộ đáng kể nhất trong AI có thể sẽ đến từ các thực thể có khả năng hiển thị cao mà các cơ quan quản lý và tòa án có thể dễ dàng xác định. Mặc dù công việc phát triển AI có thể được thực hiện bởi một người sử dụng máy tính cá nhân, tính kinh tế theo quy mô và khả năng tiếp cận nguồn vốn nhân lực và tài chính lớn hơn vẫn mang lại lợi thế đáng kể và sẽ tiếp tục làm như vậy trong tương lai. Điều này sẽ đặc biệt quan trọng nếu sức mạnh tính toán trở thành một thành phần quan trọng trong việc phát triển AI phức tạp hơn. Bộ não của con người được cho là sở hữu sức mạnh tính toán ngoại hạng, lớn hơn hai hoặc ders độ lớn hơn siêu máy tính mạnh nhất thế giới vào năm 201580 và lớn hơn tám bậc độ lớn so với máy tính xách tay điển hình hiện nay. Vào năm 2014, siêu máy tính mạnh thứ tư trên thế giới đã mất bốn mươi phút để mô phỏng một giây hoạt động não bộ của con người. Hiện tại, danh sách các nhà điều hành siêu máy tính mạnh nhất thế giới được thống trị bởi các tổ chức chính phủ, doanh nghiệp nhà nước, các tổ chức nghiên cứu lớn và các tập đoàn có vốn hóa lớn. Nếu tình trạng đó tiếp tục, thì các chính phủ quốc gia và các tập đoàn lớn, những thực thể tạo ra các nguồn rủi ro công cộng khác, có thể là những thực thể duy nhất có khả năng xây dựng các hệ thống AI mạnh trong nhiều năm. Ít nhất, các dự án được hỗ trợ bởi các thực thể như vậy sẽ có lợi thế đáng kể so với các nỗ lực xây dựng hệ thống AI tinh vi khác.

Tuy nhiên, về một khía cạnh nào đó, đầu tư ngày càng tăng của khu vực tư nhân vào AI thu hẹp phạm vi các công cụ hiệu quả mà chính phủ sử dụng. Chi tiêu lớn của khu vực tư nhân cho phát triển AI khiến cho việc nghiên cứu an toàn AI được chính phủ trợ cấp, đứng riêng lẻ, sẽ không có tác động đáng kể. Không có chi tiêu công thực sự cắt cổ, đầu tư của chính phủ vào nghiên cứu AI sẽ bị thu hẹp bởi đầu tư của khu vực tư nhân - và trừ khi có một sự kiện đại hồng thủy ở quy mô của Chiến tranh thế giới thứ hai, không có khả năng công chúng thèm muốn chi tiêu lớn của chính phủ cho các dự án AI sẽ thành hiện thực . Hơn nữa, nếu mục tiêu đầu tư vào AI của khu vực công chỉ đơn giản là nghiên cứu sự an toàn của AI và công bố thông tin về cách phát triển AI an toàn, thì vẫn cần phải có một số loại cơ chế để khuyến khích hoặc yêu cầu các nhà phát triển AI kết hợp các tính năng an toàn kết quả vào hệ thống của họ. Do đó, mặc dù nghiên cứu được chính phủ trợ cấp có thể bổ sung khuôn khổ pháp lý rộng rãi hơn cho AI, nhưng nó sẽ không phải là phản ứng pháp lý đầy đủ đối với những rủi ro công cộng mà AI sẽ tạo ra.

May mắn thay, các tổ chức pháp lý và quản lý của thế giới công nghiệp hóa cung cấp một bộ công cụ rộng và sâu, cung cấp nhiều phương pháp tiềm năng để tác động đến sự phát triển và hoạt động của AI. Ngay cả khi một khía cạnh nào đó của AI không dễ bị ảnh hưởng bởi các quy định trực tiếp từ cơ quan hành chính, nó có thể đáp ứng các khuyến khích gián tiếp do luật cấm cung cấp. Các cơ quan lập pháp, cơ quan và tòa án đều đưa ra các cơ chế có thể giúp định hướng sự phát triển của AI theo những cách có lợi về mặt kinh tế và xã hội. Phần III sẽ đề cập đến năng lực so sánh của từng loại tổ chức chính phủ này để quản lý các rủi ro công liên quan đến AI.