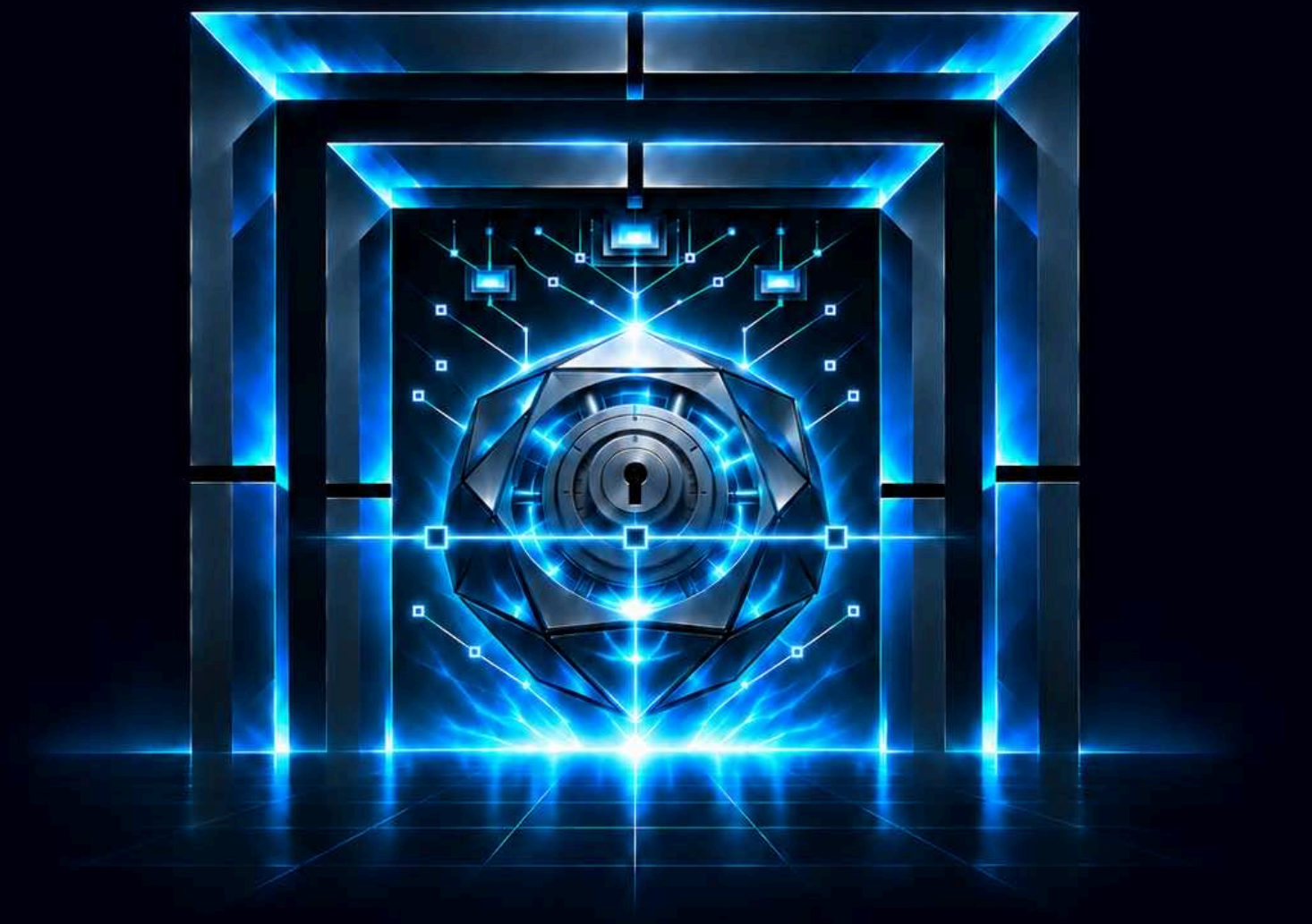
M.I.A.M.

KIỀM CHẾ LẪN NHAU TRƯỚC LEO THANG QUA TRUNG GIAN TRÍ TUỆ NHÂN TẠO

CHIẾN LƯỢC TRÍ TUỆ KỶ VẬT



TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG

HIGHLAND REFLEXIVE PRESS

HRP–RN–013–VI · ẤN BẢN NGHIÊN CỨU CÔNG KHAI

Học thuyết M.I.A.M.

Kiềm chế lẫn nhau trước leo thang qua trung gian trí tuệ nhân tạo — Chiến lược Trí tuệ Kỷ vật

TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG

HIGHLAND REFLEXIVE PRESS

Bản cập nhật ngày 15 tháng 11 năm 2026 · DOI 10.5281/zenodo.24215109
Copyright © 2026 - Tony T. Nguyen | Nguyễn Từ Vĩnh Trường - Highland Reflexive Press - All rights reserved.

Tóm tắt

Năm 2025, Dan Hendrycks, Eric Schmidt và Alexandr Wang đề xuất một chiến lược cho thời đại siêu trí tuệ dựng trên ba trụ cột: răn đe bằng phá hoại lẫn nhau (MAIM), ngăn phổ biến, và năng lực cạnh tranh, trong đó có việc tích hợp trí tuệ nhân tạo vào chỉ huy và kiểm soát quân sự (Hendrycks, Schmidt & Wang, 2025). Học thuyết M.I.A.M đi song song với cấu trúc ấy và đặt vào đó ba trụ cột đối lại.

> 1. Kiềm chế lẫn nhau thay cho phá hoại lẫn nhau. Trạng thái mặc định không phải là sẵn sàng đánh, mà là con số không: không một quả nào được khai hỏa, kể cả để trả đũa.

> 2. Không ủy quyền cho máy thay cho ngăn phổ biến. Thứ cần ngăn không chỉ là vũ khí lọt vào tay kẻ xấu, mà là quyền quyết định lọt khỏi tay con người.

> 3. Sức mạnh của sự chậm thay cho cuộc đua tốc độ. Quốc gia an toàn nhất không phải quốc gia nhanh nhất, mà là quốc gia có cấu trúc buộc mình phải chậm lại trước chiếc nút.

Hai học thuyết cùng nói về con số không, nhưng là hai con số không khác nhau. Con số không của MAIM là con số không của sự tê liệt: không ai vượt lên được, vì dự án nào nhô lên sẽ bị đối thủ làm cho tê liệt. Con số không của M.I.A.M là không có chuyện gì xảy ra: chiếc nút không bao giờ được bấm. Gốc của con số không thứ hai nằm ở một ý rất cũ của đạo Phật, thuyết vô ngã: khi không còn cái "ta" phải bảo vệ hay báo thù, thì cũng không còn ai cần đến chiếc nút.

MAIM tìm cách ngăn một quốc gia độc chiếm siêu trí tuệ. M.I.A.M được thiết kế để ngăn một nỗ lực chống độc quyền siêu trí tuệ, hoặc một cảnh báo sai về nguy cơ ấy, leo thang thành thảm họa hạt nhân.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Lời mở

Có những đầu đạn hạt nhân sinh ra để làm rộn ràng phòng làm việc của các nguyên thủ quốc gia; lại có những đầu đạn khác cũng sinh ra, nhưng gánh một vai trò hoàn toàn khác. Sứ mệnh ấy là gì, xin quý độc giả kiên nhẫn chờ tới Kịch bản thứ năm, hạ hồi sẽ phân giải.

Và mùa này, ngay giữa những căn phòng rộn ràng ấy, người ta bắt đầu rỉ tai nhau rằng tận miền hoang mạc Texas xa xôi cỏ cháy, một kẻ ban ngày mài giũa móng tay lại cả gan đặt bút viết hẳn bản học thuyết về chiếc nút bấm hạt nhân. Không ít bậc thức giả đã nhướng mày đầy vẻ bề trên—thái độ trịch thượng quen thuộc mỗi khi một kẻ ngoại đạo bước nhầm vào sảnh tiệc vốn chỉ dành riêng cho giới lễ phục hàn lâm.

Bài luận này xin phép đáp lễ bằng tất cả sự nhã nhặn tối thiểu: tự bao giờ mùi acetone và cây giũa móng lại có quyền tước đoạt tư cách lo âu cho vận mệnh nhân loại? Chẳng lẽ chỉ những bộ óc đội tóc giả trong phòng khánh tiết mới được phép bận tâm tới thời khắc địa cầu khép lại bằng một tiếng nổ khô khốc—khoảnh khắc vĩnh viễn chẳng còn ai sống sót để ghi chép, bình duyệt hay tranh cãi xem phần đúng thuộc về bên nào?

Giới tinh hoa có thể hao tâm tổn trí bàn luận xem ngón tay quyền lực nào sẽ ấn xuống nút bấm định mệnh ấy; kẻ viết những dòng này thì tuyệt nhiên không. Bởi lẽ, một vị nguyên thủ đang nóng giận hay một cỗ máy vô tri lên cơn hoang tưởng, suy cho cùng, cũng chỉ ban phát cùng một thứ ân huệ như nhau.

Điều duy nhất đáng để các bậc thông thái khắc cốt ghi tâm chính là: một khi chiếc nút đã lún xuống, cả nhân loại sẽ cùng dắt tay nhau bước vào cõi hư vô trong cuộc diệt vong bình đẳng tuyệt đối. Ở nơi đó, chẳng còn sinh linh nào sống sót để đứng lại quy kết trách nhiệm —và may mắn thay, cũng chẳng còn ban biên tập nào thức đến nửa đêm để gửi đi thêm một lá thư từ chối lịch thiệp—Regrettably!!!

Học thuyết M.I.A.M chỉ theo đuổi một kết quả: dù ai hung hăng đến đâu, số quả vũ khí hạt nhân được khai hỏa vẫn phải bằng không. Đây không phải là một mục tiêu trong nhiều mục tiêu có thể đánh đổi với nhau, bởi trong thế giới hạt nhân, chỉ cần một quả nổ là cả chùm nổ theo, và sau đó sẽ không còn một cơ quan nào đủ nguyên vẹn để rút kinh nghiệm.

Suốt thời đại hạt nhân, nhân loại được cứu không phải nhờ máy móc chính xác, mà nhờ những con người đã nghi ngờ máy móc và còn đủ thời gian để chần chừ. Trí tuệ nhân tạo đang lấy đi cả hai thứ ấy cùng một lúc: nó khiến cảnh báo trông đáng tin hơn, và nó làm cho thời gian để kiểm chứng cảnh báo ngắn đi hoặc mất luôn. Các cường quốc đã bắt đầu nói rằng con người phải nắm quyết định cuối cùng, nhưng chưa ai nói rõ đại diện ấy phải cân nhắc trong những điều kiện nào, và liệu quyết định của họ có thể định luôn số phận của cả quả địa cầu, của tám tỷ con người và muôn loài muông thú hay không.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Học thuyết này trả lời câu hỏi đó. Nó gồm một mệnh đề trung tâm, sáu nguyên tắc và một kiến trúc ba tầng khóa, được xây dựng từ những bài học mà lịch sử đã phải trả giá bằng những lần suýt mất tất cả.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Phần một. Chiếc nút là chuyện an ninh của cả loài người

Trí tuệ nhân tạo đã trở thành chuyện an ninh quốc gia. M.I.A.M đi thêm một bước: chiếc nút hạt nhân là chuyện an ninh của cả loài người, vì không quốc gia nào sống sót một mình sau khi nó được bấm.

Foucault nhìn thấy ở bom nguyên tử một nghịch lý của quyền lực hiện đại. Nhà nước lo cho dân mình sống: đo tuổi thọ, ngăn dịch bệnh và chuẩn bị cho ngày mai. Cũng nhà nước ấy giữ trong tay một thứ có thể khiến ngày mai biến mất đối với những người chưa bao giờ thuộc về nó. Người có quyền ra lệnh có thể đã được trao quyền để bảo vệ một quốc gia. Nhưng những người ở bên kia biên giới đã trao cho ông quyền gì? Và những người chưa kịp sinh ra đã ủy quyền cho ông lúc nào?

Đặt trí tuệ nhân tạo trước chiếc nút làm câu hỏi ấy khẩn cấp hơn. Cỗ máy có thể tính trong vài giây. Vị nguyên thủ có thể ký đúng thẩm quyền trong nước. Cả hai việc vẫn để nguyên câu hỏi về những người sẽ gánh hậu quả. Một mệnh lệnh có thể đi qua mọi cửa hợp pháp trong bộ máy của một nước, trong khi hậu quả của nó đi qua những biên giới mà thẩm quyền của người ký chưa từng vượt qua. Giữ quyết định trong tay con người mới chỉ trả lời ai có thể ra lệnh. M.I.A.M còn phải trả lời quyền ấy có thể vươn tới đâu. Đó là lý do chiếc nút là chuyện an ninh của cả loài người.

Bài của Hendrycks và cộng sự chia mối nguy thành bốn mảng: cạnh tranh giữa các quốc gia, khủng bố, mất kiểm soát và các chiến lược hiện có. Phần này đi theo đúng bốn mảng ấy nhưng nhìn chúng từ phía chiếc nút: ba lần thế giới suýt mất kiểm soát, bốn con đường trí tuệ nhân tạo đẩy tới leo thang, giới hạn của răn đe cổ điển, và những học thuyết đi trước.

Ba lần thế giới thoát nạn nhờ một con người chần chừ

Ngay cả trước khi có trí tuệ nhân tạo, những lần nhân loại đứng sát bờ vực nhất đều có chung một cấu trúc: một cảnh báo trông như thật, một khung thời gian bị nén lại, và một con người cụ thể đã không làm theo điều mà hệ thống đang đẩy họ tới.

Ngày 27 tháng 10 năm 1962, tàu ngầm B-59. Hãy hình dung một chiếc tàu ngầm Liên Xô nằm dưới biển Caribe giữa đỉnh điểm khủng hoảng tên lửa Cuba, ắc quy gần cạn, máy lạnh đã hỏng, hơi nóng trong khoang ngột ngạt đến mức thủy thủ ngất xỉu từng người, và đã nhiều ngày không nhận được một tin tức nào từ Moskva. Trên đầu họ, các tàu khu trục Mỹ liên tục thả bom chìm, tiếng nổ dội vào vỏ thép như búa gõ vào nồi, và không ai trong khoang biết đó chỉ là bom tín hiệu ép tàu nổi lên hay chiến tranh đã thật sự bắt đầu. Trong bụng con tàu có một ngư lôi hạt nhân mà phía Mỹ không hề biết, và việc phóng nó đòi hỏi cả ba sĩ quan cấp cao phải đồng ý. Thuyền trưởng, theo lời kể của người có mặt, đã muốn bắn và chấp nhận

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

cùng chết với đối phương, chính ủy đứng về phía ông, và chỉ còn một chữ ký. Vasily Arkhipov đã không ký. Ông thuyết phục con tàu nổi lên giữa vòng vây để chờ lệnh, và khi nắp tàu mở ra, thứ đón họ là không khí của một thế giới vẫn còn nguyên, chứ không phải bóng mây hình nấm.

Chiếc tàu ngầm ấy không phải ngòi nổ duy nhất của ngày hôm đó. Washington về sau gọi ngày 27 tháng 10 năm 1962 là Thứ Bảy Đen: cùng ngày, lực lượng chiến lược Mỹ đã ở mức báo động DEFCON 2, một máy bay do thám U-2 bị bắn rơi trên Cuba, và một chiếc U-2 khác lạc vào không phận Liên Xô khiến cả hai phía cùng cho máy bay tiêm kích cất cánh. Trong bối cảnh ấy, những người ngồi dưới đáy biển không có cách nào biết rằng phía Mỹ chỉ đang dùng loại bom chìm luyện tập để ra hiệu, cũng không có cách nào hỏi lại Moskva. Ngư lôi trên tàu là loại T-5 mang đầu đạn hạt nhân, thường được mô tả là có sức công phá cỡ quả bom Hiroshima. Phần lớn những gì chúng ta biết về không khí trong khoang đến từ lời kể của sĩ quan tình báo vô tuyến Vadim Orlov, công bố năm 2002; một số cựu thủy thủ khác kể lại nhẹ nhàng hơn, nên những chi tiết kịch tính nhất cần được đọc như lời chứng của một người, không phải biên bản.

Điều đáng nhớ nhất của câu chuyện không nằm ở cá tính của Arkhipov, mà nằm ở một quy tắc: việc phóng ngư lôi đòi hỏi sự đồng thuận của ba người, và chỉ cần một người nói không là đủ. Nhưng quy tắc ấy đã cứu thế giới nhờ một sự tình cờ: Arkhipov là tham mưu trưởng của cả đội tàu ngầm và chỉ tình cờ đi trên đúng chiếc B-59. Trên một chiếc tàu khác, ghế thứ ba có thể đã thuộc về một người dễ gật đầu hơn. M.I.A.M học từ đây một điều giản dị: quyền nói không không được phép phụ thuộc vào việc ai tình cờ ngồi ở ghế nào, mà phải được đặt sẵn vào cấu trúc, ở những chỗ mà lợi ích của người giữ khóa không trùng với lợi ích của người muốn bấm.

Ngày 26 tháng 9 năm 1983, hầm Serpukhov-15. Vừa qua nửa đêm, còi báo động trong hầm chỉ huy rú lên, màn hình lớn sáng chữ đỏ, và hệ thống vệ tinh cảnh báo sớm Oko báo rằng Mỹ đã phóng một tên lửa, rồi hai, rồi ba, rồi tới năm tên lửa đạn đạo đang bay về phía Liên Xô. Sĩ quan trực ban Stanislav Petrov ngồi giữa tiếng còi với chiếc điện thoại trong tay, biết rằng chỉ cần ông báo lên đó là một cuộc tấn công thì cỗ máy trả đũa sẽ bắt đầu quay và sẽ không còn ai kịp dừng nó lại. Ông báo lên rằng hệ thống trục trặc. Ông lập luận rằng không ai mở màn chiến tranh hạt nhân bằng năm quả tên lửa, rằng hệ thống còn mới và chưa đáng tin, và rằng radar mặt đất không xác nhận, nhưng ông không thể chứng minh mình đúng, và những phút sau đó là những phút một con người phải ngồi chờ xem trái đất có còn hay không. Không có gì rơi xuống. Điều tra về sau cho thấy vệ tinh đã nhầm ánh nắng phản chiếu trên mây thành vệt khói tên lửa.

Mùa thu năm 1983 là mùa tệ nhất để một cỗ máy báo động sai. Ba tuần trước đêm ấy, phòng không Liên Xô đã bắn rơi chuyến bay KAL 007 của Hàn Quốc, giết cả 269 người trên máy bay vì tưởng đó là máy bay do thám Mỹ. Vài tuần sau đêm ấy, NATO tiến hành cuộc tập trận Able Archer 83 diễn tập quy trình dùng vũ khí hạt nhân, và một phần giới lãnh đạo Liên Xô lo sợ đó là vỏ bọc cho một cuộc tấn công thật. Trong bầu không khí ngờ vực ấy, hệ thống vệ tinh Oko, mới đưa vào vận hành, báo về năm quả tên lửa Minuteman. Sau này người ta tìm ra nguyên nhân: một vệ tinh đã bắt nhầm ánh nắng phản chiếu trên đỉnh mây thành vệt lửa của tên lửa đang phóng.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Hãy để ý cách Petrov lập luận. Ông không có dữ liệu nào tốt hơn cỗ máy; ông chỉ có một chỗ đứng. Ông biết hệ thống còn non, biết radar mặt đất chưa xác nhận, và biết rằng một cuộc tấn công phủ đầu thật sẽ không mở màn bằng năm quả tên lửa. Một hệ thống trí tuệ nhân tạo hôm nay sẽ làm công việc của Oko tốt hơn nhiều, nhưng nó cũng sẽ làm một việc mà Oko không làm nổi: tự tin hơn. Nó sẽ gắn cho năm vệt sáng một điểm tin cậy cao, kèm một lời giải thích trôi chảy, và có thể còn đề xuất sẵn một phương án đáp trả. Điều cứu thế giới năm 1983 là một con người dám nghi ngờ một cỗ máy còn vụng về; điều đáng sợ của thập niên này là những cỗ máy không còn vụng về nữa. Câu chuyện còn một hậu kiện đáng ngẫm: vụ việc bị giữ kín suốt nhiều năm, và chính Petrov kể rằng các nhà điều tra đã tìm cách đổ lỗi cho con người thay vì thừa nhận phần cứng đã bị đưa vào vận hành quá vội. Một hệ thống trách phạt người dám nghi ngờ máy là một hệ thống đang dạy những Petrov tương lai đừng nghi ngờ nữa. Đó là lý do Nguyên tắc 6 ghi lại trách nhiệm của người chọn, nhưng không bao giờ trừng phạt người chọn dừng lại.

Ngày 25 tháng 1 năm 1995, chiếc va li hạt nhân. Chiến tranh Lạnh đã kết thúc ba năm, người ta tưởng nỗi sợ đã được cất vào bảo tàng, thì một buổi sáng một tên lửa nghiên cứu khí quyển của Na Uy rời bệ phóng, và radar Nga thấy nó giống hệt một tên lửa phóng từ tàu ngầm Mỹ để gây nổ trên cao, làm mù radar trước một cuộc tấn công thật. Na Uy đã thông báo trước về vụ phóng, nhưng thông báo đó không bao giờ tới được những người đang nhìn vào màn hình. Lần duy nhất được biết đến công khai trong lịch sử, chiếc va li hạt nhân được kích hoạt và đặt trước mặt Tổng thống Yeltsin, và một người đàn ông cùng các tướng lĩnh của mình đứng trước một quyết định chỉ có vài phút để đưa ra. Thứ cứu thế giới hôm ấy không phải trí khôn của ai, mà là vài phút theo dõi quỹ đạo, đủ để thấy chấm sáng trên màn hình đang bay ra biển chứ không bay về Moskva, và chiếc va li được đóng lại.

Tên lửa ấy là một quả Black Brant XII bốn tầng, phóng từ bãi Andøya để nghiên cứu cực quang. Radar cảnh báo sớm ở Olenegorsk phát hiện nó, và với những người trực, tốc độ và quỹ đạo của nó giống một tên lửa Trident phóng từ tàu ngầm Mỹ; khi các tầng tách ra, hình ảnh trên radar còn có thể trông như một tên lửa đang nhả nhiều đầu đạn. Na Uy đã làm đúng mọi thủ tục: thông báo đã được gửi đến Bộ Ngoại giao Nga từ trước, nhưng nó dừng lại ở đó và không bao giờ tới được quân đội. Theo một ghi chép của Bộ Tư lệnh châu Âu của Mỹ, học thuyết Nga khi ấy cho tổng thống khoảng mười phút kể từ lúc phát hiện để quyết định. Tên lửa bay khoảng hai mươi bốn phút rồi rơi xuống biển gần Spitsbergen đúng như kế hoạch, và hôm sau chính Yeltsin kể với báo chí rằng lần đầu tiên ông đã phải dùng tới chiếc va li nhỏ có nút bấm luôn theo mình.

Nếu B-59 là bài học về quy tắc và Petrov là bài học về sự nghi ngờ, thì năm 1995 là bài học về đối thoại. Không ai trong câu chuyện này có ác ý. Các nhà khoa học Na Uy đã báo trước, người trực radar đã làm đúng phận sự, và chiếc va li đã được mở đúng quy trình. Thứ hỏng là một sợi dây thông tin đứt giữa đường: một lá thư đã gửi mà không ai xác nhận đã nhận. Đó là lý do M.I.A.M không chấp nhận một thông báo một chiều làm bằng chứng đối thoại, mà đòi mỗi lượt nói phải được bên kia xác nhận đã nhận (Nguyên tắc 2). Mười phút của năm 1995 đã đủ, vì đó là một quả tên lửa khoa học bay ra biển và một quỹ đạo ai cũng đọc được. Không có gì bảo đảm lần sau mọi thứ cũng rõ ràng như vậy.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Ba câu chuyện để lại ba bài học mà học thuyết này biến thành nguyên tắc. Bài học thứ nhất là máy móc có thể sai một cách rất thuyết phục. Bài học thứ hai là thời gian chính là thứ đã cứu thế giới, đó là vài phút mà Petrov chờ xác nhận và vài phút mà Moskva theo dõi quỹ đạo. Bài học thứ ba là quy tắc đòi hỏi nhiều người cùng đồng ý, như trên B-59, đã cho một con người quyền nói không. Chính bởi sự sống còn của nhân loại đã từng phụ thuộc vào may mắn và cá tính của vài người, học thuyết này đòi hỏi biến sự chần chừ của họ thành cấu trúc, để lần sau không còn phải trông chờ vào may mắn.

Và trước cả ba lần ấy, một ổ khóa đã không mở, trên đất Việt. Mùa xuân năm 1954, khi quân Pháp bị vây ở Điện Biên Phủ, Hoa Kỳ cân nhắc một phương án can thiệp bằng không quân mang tên Chiến dịch Kền Kền (Operation Vulture). Người ủng hộ mạnh nhất là Đô đốc Arthur Radford, chủ tịch Hội đồng Tham mưu trưởng Liên quân, dựa trên một nghiên cứu cho rằng ba quả bom nguyên tử chiến thuật là đủ để phá vòng vây. Tổng thống Eisenhower không bác bỏ thẳng; ông đặt điều kiện: Quốc hội phải chấp thuận trước, và nước Anh phải cùng tham gia. Nước Anh không tham gia, các nhà lập pháp không muốn một cuộc can thiệp đơn phương, và ngày 7 tháng 5 năm 1954 cứ điểm thất thủ mà không có quả bom nào rơi xuống.

Câu chuyện này khác ba câu chuyện trên ở một điểm quan trọng. Ở B-59, ở Serpukhov-15, ở Moskva năm 1995, thế giới được cứu nhờ một con người chần chừ. Ở Điện Biên Phủ, nó được cứu nhờ một cấu trúc: một quyết định hạt nhân bị buộc phải đi qua những bên có lợi ích khác nhau, và chỉ cần một bên không đồng ý là đủ. Đó chính là lớp khóa thứ hai của M.I.A.M, đã từng vận hành dù chưa ai gọi tên nó. Và nó vận hành trên chính mảnh đất mà người viết học thuyết này gọi là quê hương: dân tộc suýt làm mục tiêu của ba quả bom năm 1954 nay có một người con viết ra học thuyết này với mong muốn rằng không dân tộc nào khác có khả năng nhận lấy viễn cảnh đó.

Chẩn đoán: bốn con đường AI đẩy con người tới leo thang

Mối nguy không nằm ở chỗ AI tự tay bấm nút, vì điều đó các cường quốc đã tuyên bố sẽ không cho phép. Mối nguy nằm ở chỗ AI đứng giữa thực tại và con người, định hình những gì con người nhìn thấy, tin vào và kịp cân nhắc. Như một cựu tướng Mỹ đã cảnh báo, rủi ro của việc rút ngắn thời gian quyết định hạt nhân nhiều khả năng lớn hơn lợi ích. Học thuyết này nhận diện bốn con đường cụ thể.

> 1. Nén thời gian. Với chính sách phóng khi có cảnh báo, người ra quyết định có thể chỉ còn chưa tới 20 phút. Mỗi cải tiến về tốc độ phát hiện lại tạo ra áp lực để hành động nhanh hơn, chứ hiếm khi được dùng để cho con người thêm thời gian.

> 2. Độ tin cậy giả. Một cảnh báo kèm theo một con số xác suất trông khoa học hơn lời nói của một con người. Hiện tượng thiên kiến tự động hóa khiến người ta có xu hướng tin máy hơn tin chính sự nghi ngờ của mình, đúng kiểu nghi ngờ đã cứu Petrov.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

> 3. Đóng khung lựa chọn. Khi máy chỉ đưa ra một phương án, hoặc đưa ra ba phương án mà hai cái chỉ để làm nền, con người ký vào đó vẫn mang danh là người quyết định, nhưng quyết định thật đã được đưa ra từ trước.

> 4. Dây chuyền phản ứng giữa hai bên. Khi cả hai phía đều dùng máy để đọc đối phương, một hành động phòng thủ của bên này có thể bị máy của bên kia đọc thành dấu hiệu tấn công, và ngược lại. Leo thang có thể hình thành từ hai cỗ máy phản ứng với nhau, trước khi có bất kỳ con người nào thật sự muốn chiến tranh.

Vịnh Bắc Bộ, tháng 8 năm 1964, cho thấy con đường thứ nhất trong bốn con đường trên có thể đi xa tới đâu, ngay cả khi chưa có trí tuệ nhân tạo. Đêm 4 tháng 8, hai tàu khu trục Mỹ báo bị tàu phóng lôi tấn công, dựa trên những tín hiệu sonar và radar trong thời tiết xấu. Chính chỉ huy hải đội, John Herrick, đã gửi điện bày tỏ nghi ngờ về nhiều mục tiêu phát hiện được và đề nghị tìm xác tàu vào sáng hôm sau. Vậy mà chỉ ba ngày sau, Quốc hội Mỹ đã thông qua nghị quyết trao quyền mở rộng chiến tranh. Năm 2005, Cơ quan An ninh Quốc gia Mỹ giải mật nghiên cứu của sử gia Robert Hanyok, xác nhận không hề có cuộc tấn công thứ hai và tình báo tín hiệu khi đó đã bị trình bày lệch (Hanyok, 2001). Không quả bom hạt nhân nào được dùng, nhưng đó là bài học rõ nhất về cách một bằng chứng mơ hồ, đi nhanh hơn sự nghi ngờ, có thể trở thành nền móng của một cuộc chiến kéo dài hơn một thập kỷ.

Bốn con đường này có một điểm chung: không con đường nào bị chặn bởi câu "con người phải quyết định". Chúng chỉ bị chặn khi các điều kiện của quyết định được quy định rõ và được thực thi bằng cấu trúc.

Đọc lại chương ba của Hendrycks từ chỗ chiếc nút

Chương ba của bài "Superintelligence Strategy" lập luận rằng trí tuệ nhân tạo đe dọa an ninh quốc gia qua ba ngả: cạnh tranh giữa các cường quốc, khủng bố, và mất kiểm soát. Học thuyết M.I.A.M đồng ý với cả ba chẩn đoán. Điều nó không đồng ý là phương thuốc, và chỗ bất đồng hiện rõ nhất khi đặt từng chẩn đoán cạnh chiếc nút hạt nhân.

Cạnh tranh chiến lược. Hendrycks và cộng sự lo rằng trí tuệ nhân tạo có thể sinh ra những siêu vũ khí làm lung lay thế răn đe hạt nhân, và nỗi lo bị đối thủ vượt mặt có thể đẩy các nhà lãnh đạo tới hành động phòng ngừa (Hendrycks, Schmidt & Wang, 2025, mục 3.1.2). Chẩn đoán ấy đúng, nhưng nó chỉ ra một điều mà chính bài viết không đi tới cùng: mối nguy không nằm ở siêu vũ khí, mà nằm ở cái đồng hồ. Một nhà lãnh đạo tin rằng đối thủ sắp có lợi thế quyết định sẽ thấy thời gian đang chống lại mình, và người thấy thời gian chống lại mình là người dễ ra tay trước nhất. MAIM trả lời nỗi sợ ấy bằng một nỗi sợ khác, là bị phá hoại. M.I.A.M trả lời bằng cách lấy lại thời gian, để không quyết định nào được phép sinh ra từ cảm giác đang hết giờ.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Khủng bố và tấn công hạ tầng. Bài viết chỉ ra rằng các cuộc tấn công mạng có trí tuệ nhân tạo trợ giúp sẽ khó truy ra thủ phạm hơn, và trong cán cân công thủ, kẻ tấn công chỉ cần tìm ra một lỗ hổng trong khi người phòng thủ phải vá tất cả (mục 3.2.2 và 3.2.3). Nếu cả hai điều ấy đúng, thì hệ quả logic là không quyết định trả đũa nào được dựa trên việc máy móc quy trách nhiệm cho một thủ phạm, vì chính việc quy trách nhiệm đã trở nên kém tin cậy. M.I.A.M còn đảo ngược cán cân công thủ ở đúng chỗ quan trọng nhất. Với một hệ thống mặc định là dừng, kẻ tấn công tìm được một lỗ hổng chỉ thu về một trạng thái dừng; muốn có một quyết định phóng, hắn phải mở cùng lúc mọi ổ khóa, vốn nằm trong tay những bên có lợi ích đối nghịch nhau. Ở chiếc nút, chính kẻ tấn công mới là bên phải thắng ở mọi điểm.

Mất kiểm soát. Đây là chỗ hai học thuyết gần nhau nhất. Hendrycks và cộng sự mô tả rất chính xác cách con người mất quyền kiểm soát: không phải qua một cuộc đảo chính kịch tính của máy móc, mà qua một chuỗi những quyết định nhỏ, quyết định nào cũng có vẻ hợp lý, cho tới khi con người không còn đủ khả năng điều khiển những quyết định then chốt (mục 3.3.1). Họ cũng kết luận rằng kiểm soát phải là một quá trình liên tục chứ không phải một lần sửa là xong (mục 3.3.3). M.I.A.M đồng ý trọn vẹn, và chỉ hỏi một câu: nếu sự xói mòn kiểm soát diễn ra qua những bước nhỏ hợp lý, thì việc tích hợp trí tuệ nhân tạo vào chỉ huy và kiểm soát, mà chính bài viết khuyến nghị ở chương sáu, có phải là một trong những bước nhỏ hợp lý ấy không? Chương ba đã trả lời câu hỏi đó trước khi chương sáu kịp đặt ra nó. M.I.A.M chính là quá trình kiểm soát liên tục mà chương ba đòi hỏi, đặt ở nơi duy nhất mà một lần mất kiểm soát là không bao giờ lấy lại được. Nghe thì có vẻ trừu tượng, nhưng chuỗi bước nhỏ ấy không phải chuyện giả định. Nó đã từng xảy ra với chính vũ khí hạt nhân, và xảy ra trên đất Việt, ở một căn cứ tên là Khe Sanh.

Khe Sanh chứng minh một điều: vũ khí hạt nhân có thể tiến sát tới chỗ được dùng mà không cần ai quyết định dùng nó, chỉ qua một chuỗi bước chuẩn bị nhỏ, bước nào cũng có vẻ hợp lý. Đó đúng là kiểu mất kiểm soát mà chính Hendrycks và cộng sự cảnh báo ở chương ba, và nó đã xảy ra thật, trên đất Việt. Đầu năm 1968, khoảng sáu nghìn lính thủy quân lục chiến Mỹ bị vây ở căn cứ Khe Sanh, một cao nguyên hẻo lánh gần vĩ tuyến 17 và biên giới Lào. Báo chí Mỹ khi ấy so sánh nó với Điện Biên Phủ năm 1954, và nỗi sợ trong giới chỉ huy là một thất bại y hệt: một căn cứ bị cô lập, bị pháo kích ngày đêm, có thể thất thủ.

Trong bối cảnh đó, Tướng William Westmoreland, tư lệnh quân Mỹ ở Việt Nam, phê duyệt một kế hoạch dự phòng mang mật danh Fracture Jaw. Kế hoạch không ra lệnh ném bom nguyên tử; nó chỉ chuẩn bị đưa vũ khí hạt nhân vào miền Nam Việt Nam để, nếu Khe Sanh sắp mất, chúng có thể được dùng trong thời gian ngắn. Từng bước một, kế hoạch tiến khá xa ở Bộ Tư lệnh Thái Bình Dương: lập phương án, bàn chuyện vận chuyển, chọn cách triển khai. Chỉ khi tin tức tới tai Nhà Trắng, Tổng thống Johnson mới ra lệnh dừng lại. Các sử gia còn tranh luận Washington biết trước tới đâu, vì chính Washington cũng từng hỏi về khả năng dùng vũ khí hạt nhân chiến thuật.

Điều đáng sợ của Khe Sanh không nằm ở một vị tướng điên rồ. Không ai trong câu chuyện này quyết định dùng bom nguyên tử. Mỗi người chỉ làm một việc có vẻ hợp lý trong phạm vi của mình: một kế hoạch dự phòng, một bước chuẩn bị, một phương án hậu cần. Cộng lại, những bước nhỏ ấy đưa vũ khí hạt nhân tới

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

gần một chiến trường hơn bao giờ hết, trước khi người có quyền cao nhất kịp biết. Đó chính là kiểu mất kiểm soát mà Hendrycks và cộng sự mô tả, và cũng là lý do M.I.A.M đặt ổ khóa ở chính quyết định cuối cùng, chứ không trông chờ ai đó kịp phát hiện giữa chừng.

Vì vậy, Khe Sanh đặt ra hai câu hỏi chứ không phải một. Ai có quyền dừng một việc khi biết nó đang diễn ra? Và điều gì buộc bộ máy phải cho người ấy biết đủ sớm để lời từ chối còn có nghĩa? Một chữ ký ở cuối quy trình có thể trả lời câu thứ nhất. Câu thứ hai đòi hỏi người có thẩm quyền vẫn còn thời gian xem lại căn cứ, còn thấy được những đánh giá trái chiều, và còn có thể nói "chưa đủ căn cứ" với hiệu lực thực sự. Thẩm quyền con người phải được giữ từ lúc tình thế bắt đầu được chọn lọc, đóng khung và trình bày, chứ không chỉ ở cửa phê duyệt cuối cùng. Nguyên tắc 1 và 6 bảo vệ sự đa dạng của phương án, quyền bác bỏ cách đóng khung của máy và trách nhiệm đối với lựa chọn; Nguyên tắc 2 và 3 bảo vệ khoảng thời gian để phản biện và kiểm chứng còn có ý nghĩa. Những ổ khóa không chỉ phủ quyết bước cuối. Chúng bảo toàn các điều kiện để lời từ chối của con người vẫn là một quyết định, thay vì một chữ ký đặt lên con đường mà hệ thống đã chọn sẵn.

Con số không của răn đe cổ điển

Từ khi Bernard Brodie viết rằng từ nay mục đích của quân đội là ngăn chiến tranh chứ không phải thắng chiến tranh (Brodie, 1946), học thuyết răn đe đã đứng trên một lời hứa: vì ai bắn trước cũng sẽ bị hủy diệt, nên sẽ không ai bắn. Thomas Schelling biến lời hứa ấy thành một môn khoa học về sức mạnh của sự đe dọa (Schelling, 1960; 1966), và phải công bằng mà nói, học thuyết này có một thành tích khó cãi: từ năm 1945 tới nay, không một quả vũ khí hạt nhân nào được khai hỏa trong chiến tranh.

Nhưng con số không ấy là con số không của ai? Ngay từ năm 1959, Albert Wohlstetter đã cảnh báo rằng thế cân bằng khủng bố mỏng manh hơn người ta tưởng nhiều (Wohlstetter, 1959), và Scott Sagan về sau lật lại hồ sơ để chỉ ra rằng chính những hệ thống được thiết kế để răn đe đã nhiều lần suýt gây ra tai nạn (Sagan, 1993). Ba câu chuyện ở Phần một là ba trang trong cuốn sổ ấy. Trên tàu B-59, trong hầm Serpukhov-15 và trước chiếc va li năm 1995, học thuyết răn đe không làm gì để giữ con số không; trái lại, chính logic của nó, bắn trả cho kịp trước khi bị diệt, là thứ đã đẩy những con người ấy tới sát nút bấm. Răn đe xứng đáng được ghi công vì đã giữ các cường quốc khỏi một cuộc chiến tranh lớn suốt nhiều thập kỷ, nhưng ở những đêm sát nút bấm nhất, con số không được giữ lại nhờ may mắn và nhờ vài con người dám chần chừ ngược với chính logic của nó.

Gần tám mươi năm qua, người ta ghi công cho răn đe vì thế giới chưa nổ tung. Có thể răn đe đã ngăn được những cuộc chiến cố ý, khi một nhà lãnh đạo tính toán rồi chủ động ra tay; điều đó không ai chứng minh được, mà cũng không ai bác bỏ được. Nhưng có một loại hiểm họa mà răn đe hoàn toàn bất lực: chiến tranh do nhầm lẫn. Với loại hiểm họa này, chính logic của răn đe, phải bắn trả cho kịp trước khi bị diệt, là thứ đẩy con người tới sát nút bấm. Và trong mọi lần thế giới suýt nổ vì nhầm lẫn mà hồ sơ công khai còn ghi lại, thứ giữ con số không chưa bao giờ là răn đe, mà là một con người chịu chần chừ ngược với chính logic ấy.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Năm lần vũ khí hạt nhân tiến sát chỗ dùng — và thứ đã chặn lại

1954	**Điện Biên Phủ**	Chiến dịch Kền Kền cân nhắc ba quả bom nguyên tử chiến thuật.	**Anh từ chối tham gia và Quốc hội Mỹ không ủng hộ đơn phương.**
1962	**Tàu ngầm B-59**	Ngư lôi hạt nhân chỉ còn thiếu một sự đồng thuận để phóng.	**Vasily Arkhipov không đồng ý.**
1968	**Khe Sanh**	Fracture Jaw chuẩn bị phương án hạt nhân cho cuộc vây hãm.	**Tổng thống Johnson ra lệnh dừng kế hoạch.**
1983	**Serpukhov-15**	Oko báo năm tên lửa Minuteman đang bay tới.	**Stanislav Petrov kết luận đó là báo động giả.**
1995	**Tên lửa Na Uy**	Radar Nga đọc vụ phóng khoa học thành một tên lửa Trident.	**Vài phút được dùng để theo dõi quỹ đạo.**

ĐIỀU ĐẨY TỚI CHỖ DÙNG THỨ ĐÃ CHẶN LẠI

Đặt năm lần ấy cạnh nhau, cột bên phải không có dòng nào ghi răn đe.

Năm lần ấy mới chỉ là những lần được kể nhiều nhất. Hồ sơ đã giải mật còn ghi lại nhiều lần khác, và một nghiên cứu của Chatham House năm 2014 đếm được mười ba lần, trong khoảng từ 1962 đến 2002, vũ khí hạt nhân suýt bị dùng vì tính sai, truyền tin sai hoặc lỗi kỹ thuật (Lewis và cộng sự, 2014). Bảng dưới đây chỉ chọn những lần mà một hệ thống cảnh báo đã báo về một cuộc tấn công không hề có thật.

THỜI ĐIỂM	HỆ THỐNG BÁO GÌ	NGUYÊN NHÂN THẬT	THỨ ĐÃ CHẶN LẠI
5/10/1960, radar cảnh báo sớm ở Thule, Greenland	Một đợt tấn công tên lửa lớn của Liên Xô, với độ chắc chắn 99,9%	Mặt trăng mọc phản xạ sóng radar	Con người nghi ngờ và kiểm tra lại thay vì làm theo cảnh báo
24/11/1961, Bộ Tư lệnh Không quân Chiến lược Mỹ	Mọi đường liên lạc với NORAD và ba trạm radar cảnh báo sớm cùng tắt, có thể là dấu hiệu đang bị tấn công	Một trạm chuyển tiếp duy nhất bị hỏng	Một máy bay B-52 đang bay tuần liên lạc bằng radio với Thule và xác nhận không có cuộc tấn công nào
25/10/1962, Duluth và Volk Field	Báo động có kẻ phá hoại; riêng ở Volk Field, chuông lại báo cho các máy bay đánh chặn F-106A mang vũ khí hạt nhân xuất kích	Một con gấu leo hàng rào, cộng với một đường dây báo động bị nối sai (Sagan, 1993)	Một sĩ quan gọi về Duluth hỏi lại, và một chiếc xe chạy ra đường băng chặn máy bay
9/11/1979, NORAD	Một cuộc tấn công hạt nhân toàn diện của Liên Xô	Băng diễn tập bị nạp nhầm vào hệ thống cảnh báo thật	Đối chiếu với vệ tinh và radar trong vài phút, ngay trước khi cố vấn an ninh quốc gia gọi cho tổng thống
3 và 6/6/1980, NORAD	Một đợt tấn công lớn đang bay tới	Một con chip giá 46 xu bị hỏng, cộng với lỗi thiết kế thông điệp	Các trung tâm cảnh báo so sánh nhiều nguồn khác nhau trong vài phút
26/9/1983, vệ tinh Oko	Năm tên lửa Minuteman	Nắng phản chiếu trên mây	Một trực ban không tin máy: Stanislav Petrov
25/1/1995, radar Olenegorsk	Một tên lửa Trident phóng từ tàu ngầm	Tên lửa nghiên cứu cực quang của Na Uy	Vài phút theo dõi quỹ đạo

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Đọc dọc cột thứ ba thì thấy một điều gần như buồn cười: mặt trăng, một con gấu, một cuộn băng, một con chip 46 xu, một vệt nắng, một tên lửa đo cực quang. Trong cột ấy không có một kẻ thù nào. Đọc dọc cột thứ tư thì thấy một điều quan trọng hơn: không lần nào thứ cứu thế giới là răn đe, cũng không lần nào là tốc độ. Lần nào cũng là một con người chịu nghi ngờ, một nguồn tin thứ hai để đối chiếu, và vài phút để làm việc đó. Đó chính là ba thứ mà Nguyên tắc 1, Nguyên tắc 3 và mạng lưới cảnh báo sớm chung muốn biến thành cấu trúc. Hãy nhớ thêm con số của năm 1960: 99,9%. Cỗ máy đầu tiên của thời đại tên lửa đã biết nói bằng xác suất từ trước khi có trí tuệ nhân tạo, và nó đã tự tin gần như tuyệt đối về một cuộc tấn công không có thật. Nghiên cứu của Chatham House cũng đi tới cùng kết luận: đã nhiều lần, chính phán đoán của từng cá nhân, có khi đi ngược cả quy trình, mới là thứ giữ cho vũ khí không được dùng.

M.I.A.M nhận lấy đúng con số ấy, nhưng đổi chủ cho nó. Ở răn đe, con số không là một kết quả mong đợi, sống nhờ vào việc không ai dại dột; ở M.I.A.M, con số không là một bất biến được xây dựng bằng cấu trúc, sống nhờ vào việc các ổ khóa không bao giờ mở cùng lúc. Răn đe hỏi: "Ai dám bắn trước?"; M.I.A.M hỏi: "Ai có quyền mở được tất cả các ổ khóa để cả chùm tên lửa bay lên ?", và câu trả lời là : không ai cả.

Những học thuyết đi trước khác

Với phóng khi có cảnh báo. Đây là chỗ M.I.A.M đối lập trực tiếp. Chính sách phóng khi có cảnh báo biến cảnh báo thành lý do để hành động trong vài phút, còn M.I.A.M biến cảnh báo thành lý do để chờ lâu hơn và đàm phán nhiều hơn. Cả ba lần suýt tận thế ở Phần một đều là những lần mà nếu áp dụng phóng khi có cảnh báo một cách máy móc thì thế giới đã không còn.

Với quy tắc hai người. Quy tắc đòi hỏi hai hoặc ba sĩ quan cùng đồng ý, như trên tàu B-59, là tổ tiên trực tiếp của M.I.A.M. M.I.A.M mở rộng ý tưởng ấy từ bên trong một con tàu ra ba tầng: giữa những người có thẩm quyền trong một nước, giữa hai nước đối đầu, và giữa các nước với nhân loại.

Với cam kết con người kiểm soát. Tuyên bố chung của Mỹ và Trung Quốc năm 2024, theo một cựu tướng Mỹ, phản ánh sự đồng thuận quốc tế đang lớn dần rằng AI không được thay thế phán đoán của con người trong quyết định hạt nhân. Nhưng các nhà phân tích cũng cảnh báo rằng ngay cả khi con người nắm quyết định cuối cùng, điều đó có thể vẫn chưa đủ để ngăn một lỗi của AI dẫn tới thảm họa. M.I.A.M là một câu trả lời cho khoảng thiếu hụt đó: nó biến cam kết thành các điều kiện kiểm chứng được.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Phần hai. Trụ cột thứ nhất: Kiềm chế lẫn nhau

Trụ cột thứ nhất trả lời đúng câu hỏi mà MAIM trả lời bằng sự đe dọa: làm sao để không ai bấm nút.

Mệnh đề trung tâm: bất biến không-phóng

Trong mọi hệ thống vận hành theo M.I.A.M, số quyết định phóng được xác nhận là hợp lệ luôn bằng không, trừ khi tất cả các ổ khóa cùng mở một lúc; và các ổ khóa ấy được giao cho những bên có lợi ích đối nghịch nhau đến mức chúng không bao giờ mở cùng nhau.

Mệnh đề này có hai vế. Vế thứ nhất là một điều kiện logic: bất kỳ ổ khóa nào còn đóng, vì bị từ chối, vì vắng mặt, vì mất liên lạc hay vì không kiểm chứng được, thì kết quả là không. Vế thứ hai là một lựa chọn thiết kế có chủ đích: người giữ các ổ khóa được chọn sao cho điều kiện "tất cả cùng mở" gần như không thể xảy ra trong thực tế.

Học thuyết này không tìm cách ngăn ai đó nổi giận, đe dọa hay hung hăng, vì đó là lĩnh vực của chính trị và của lương tri. Điều nó làm là lấy đi từng con đường mà cơn giận có thể đi để biến thành một mệnh lệnh hợp lệ. Một quốc gia vẫn giữ toàn quyền huấn luyện quân đội, duy trì tư thế sẵn sàng và phòng thủ bờ cõi, nhưng đúng vào lúc bàn tay chạm tới cái nút, sẽ luôn có một thứ khiến nó khựng và giật lại.

Nói một cách giản dị: số phận của hành tinh này không thể được định đoạt trong cuộc đối đầu giữa hai nhân vật, dù quyền lực của họ lớn đến đâu. Nó là gia sản chung của tám tỷ con người, và học thuyết này chính là cách để tám tỷ con người ấy, vốn chưa từng được mời vào căn phòng có chiếc nút, cất lên tiếng nói của mình.

Con số không là trạng thái mặc định, kể cả khi đã bị đánh

Học thuyết răn đe cổ điển giữ hòa bình bằng lời hứa đáp trả chắc chắn: ai đánh trước sẽ bị đánh lại. M.I.A.M từ bỏ lời hứa ấy, và nói thẳng ra điều đó. Không chỉ đòn đánh đầu, mà cả đòn trả đũa cũng nằm dưới cùng những ổ khóa.

Lý do nằm ngay trong bản chất của vũ khí hạt nhân. Một vụ nổ kéo theo cả chùm nổ, và sau đó thế giới không còn ai để bảo vệ, cũng không còn ai để rút kinh nghiệm. Đòn trả đũa không cứu được thành phố nào đã cháy; nó chỉ bảo đảm rằng thêm nhiều thành phố nữa sẽ cháy theo.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Người phản biện sẽ nói rằng không có trả đũa thì đối phương không còn gì để sợ. M.I.A.M thừa nhận cái giá ấy, và đặt sự răn đe ở chỗ khác: mọi phê duyệt đơn phương đều bị ghi vĩnh viễn vào sổ cái lịch sử, kẻ đứng ngoài mạng lưới mất quyền được bảo vệ chung, và không một quả nào có thể rời bệ phóng khi chính bên bị nhắm chưa đồng ý. Học thuyết này không hứa rằng sẽ không ai muốn đánh; kiến trúc của nó được thiết kế để ý muốn ấy không còn con đường hợp lệ nào dẫn tới một vụ phóng.

Con số không giữ lại điều gì: Souvenir Intelligence

Ngôn ngữ an ninh đếm những quốc gia còn tồn tại. Souvenir Intelligence hỏi liệu còn ai sống để nhớ điều quốc gia ấy lẽ ra phải bảo vệ hay không. Một lần phóng bằng không giữ lại nhiều hơn lãnh thổ: người ta còn sống để kể, để cãi lại bản tường trình chính thức, để thương tiếc người đã mất và để sửa sai cho thế hệ sau. Căn bếp bình thường có việc ở ngay đây. Mẹ vẫn còn đứng nấu ăn bên bếp; ngoài vườn vẫn có nắng, có gió, có tiếng chim; một gia đình vẫn còn có thể bất đồng về chuyện đã xảy ra. Đó không phải đồ trang trí quanh thước đo của học thuyết. Đó là lý do thước đo tồn tại.

Vô ngã khiến thước đo ấy làm việc ngay trong lập luận về trả đũa. Khi một lãnh đạo nói "chúng ta phải đáp trả", chữ "chúng ta" đang chỉ ai? Những người bị hại, chính quyền phải giữ trật tự công, hay hình ảnh về quốc gia mà người lãnh đạo cảm thấy mình buộc phải báo thù? Ba thứ ấy có thể tách khỏi nhau. M.I.A.M không đòi một cộng đồng quên người đã chết hay từ bỏ quyền phòng vệ. Nó hỏi việc nhân số người chết lên phục vụ người đang sống, hay phục vụ một cái ngã quốc gia bị tổn thương đã nhầm báo thù với bảo vệ. Buông cái ngã ấy trước chiếc nút không phải là thụ động. Đó là từ chối đốt thêm những căn bếp khác chỉ để chứng minh rằng nỗi đau của mình có quyền lực.

Phạm vi: chỉ một chiếc nút

M.I.A.M chỉ áp lên một loại quyết định duy nhất: quyết định khai hỏa vũ khí hạt nhân. Nó không làm chậm việc phòng thủ mạng, không trì hoãn việc ứng phó một dịch bệnh, không trói tay các lực lượng quy ước trước một cuộc tấn công, và không ngăn một quốc gia đặt mình vào tư thế phòng thủ sẵn sàng. Mọi phản ứng dưới ngưỡng hạt nhân vẫn đi theo luật lệ và học thuyết hiện hành của mỗi nước. Vì vậy, lập luận rằng đối phương có thể lợi dụng thời gian chờ để tấn công mạng hay phát tán mầm bệnh không đánh vào M.I.A.M, vì trong khoảng thời gian ấy mọi công cụ phòng vệ khác vẫn mở nguyên. Thứ duy nhất bị giữ lại là thứ duy nhất mà một khi đã dùng thì không còn ai để phòng vệ nữa.

Hai lớp khóa: lớp tự làm và lớp cùng làm

Lời phản biện mạnh nhất nhắm vào M.I.A.M là: quốc gia nào tự nguyện gắn các ổ khóa trước sẽ tự trói tay mình trong khi đối thủ đứng ngoài. Lời phản biện ấy đúng, nếu mọi ổ khóa phải được gắn cùng một lúc và gắn đơn phương. Học thuyết không đòi điều đó, vì các nguyên tắc của nó chia thành hai lớp khác nhau về bản chất.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Lớp thứ nhất là lớp một quốc gia tự làm được mà không mất gì: dự báo của máy chỉ là chứng cứ, không bao giờ là thẩm quyền (Nguyên tắc 1); cảnh báo của máy chỉ được làm chậm lại (Nguyên tắc 3); và luôn có ít nhất hai phương án, người chọn phải ký vào lựa chọn của mình (Nguyên tắc 6). Lớp này chỉ chặn những lần bấm vì lầm. Nó không đụng tới năng lực răn đe của bất kỳ ai, nên một quốc gia gắn nó trước chỉ có lợi: bớt đi nguy cơ tự hủy vì một vệt nắng trên màn hình.

Lớp thứ hai là lớp chỉ có hiệu lực khi các bên cùng vào: nghĩa vụ đối thoại hai chiều (Nguyên tắc 2), hai chìa khóa két sắt (Nguyên tắc 4) và ổ khóa của nhân loại (Nguyên tắc 5). Lớp này giống mọi hiệp ước giải trừ vũ khí từng có hiệu lực ở chỗ nó chỉ ràng buộc khi các bên cùng ký, cùng một thời điểm, và kiểm chứng được lẫn nhau. Không quốc gia nào được yêu cầu gắn nó một mình. Theo ngôn ngữ của lý thuyết trò chơi, M.I.A.M không đòi ai đặt cược trong thế lưỡng nan của người tù; nó đề nghị các bên cùng chuyển sang một trò chơi khác, nơi hợp tác chỉ bắt đầu khi sự hợp tác của bên kia đã được kiểm chứng.

Con số không tuyệt đối là đích đến của lớp thứ hai, không phải điều kiện để bắt đầu. Đó là lý do lộ trình đi từ lớp thứ nhất: một thế giới mà mọi cường quốc mới chỉ gắn lớp ấy đã an toàn hơn thế giới năm 1983, và chính sự quen thuộc với lớp ấy là nền để các bên tin nhau đủ để bước sang lớp thứ hai.

Hai lớp khóa của M.I.A.M

01 · TỰ GÁN **Lớp tự làm**	02 · CÙNG VÀO **Lớp cùng làm**
Nguyên tắc 1 · Dự báo là chứng cứ, không phải thẩm quyền. Nguyên tắc 3 · Báo động chỉ được làm chậm quy trình. Nguyên tắc 6 · Luôn có nhiều lựa chọn và người chọn phải chịu trách nhiệm.	Nguyên tắc 2 · Đối thoại hai chiều. Nguyên tắc 4 · Hai chìa khóa két sắt. Nguyên tắc 5 · Ổ khóa cuối của nhân loại.

Lớp thứ nhất chặn những lần bấm vì lầm. Lớp thứ hai chỉ ràng buộc khi mọi bên cùng ký và kiểm chứng được nhau.

Không quốc gia nào bị đòi tự trói tay một mình: lớp bên trái chỉ chặn những lần bấm vì lầm, lớp bên phải chỉ ràng buộc khi mọi bên cùng ký.

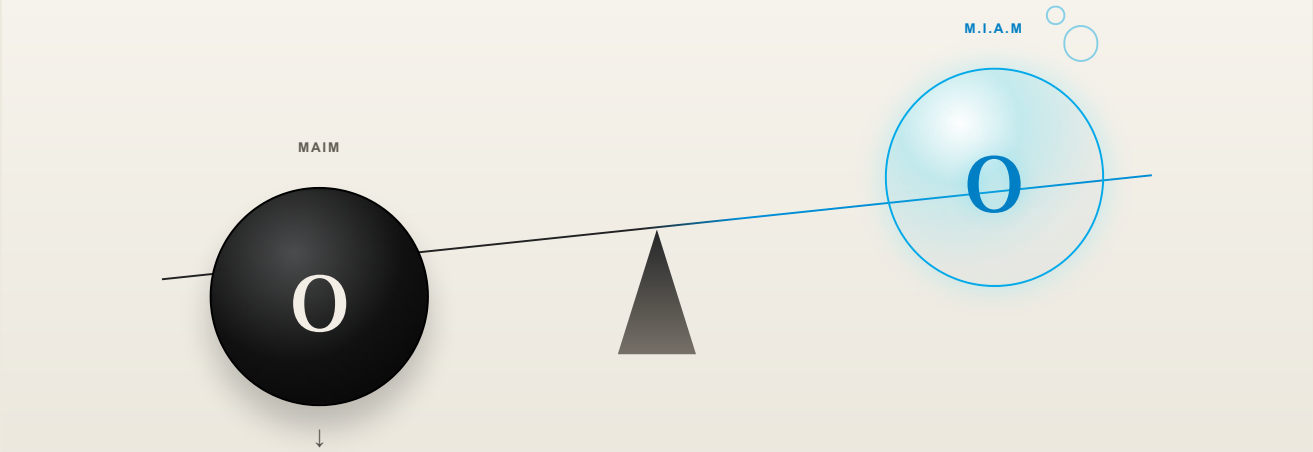
Hai con số không

Con số không của MAIM là con số không của sự tê liệt, đúng như cái tên của nó. Những người đề xướng MAIM không mong thế giới đổ nát; họ mong ngăn được một cuộc chiến. Nhưng cách họ giữ con số không là lời đe dọa làm tê liệt: dự án của đối phương có thể bị phá, trung tâm dữ liệu có thể bị đánh, và thế cân bằng được giữ bằng việc ai cũng tin rằng điều đó có thể xảy ra. Đó là con số không sinh ra từ cái ngã, khi mỗi quốc gia giữ lấy vị thế của mình bằng lời dọa làm tê liệt người khác.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Con số không của M.I.A.M là con số không của việc không có gì xảy ra. Không có đổ nát nào được hứa hẹn, chỉ có một chiếc nút mãi mãi không được bấm. Đạo Phật gọi gốc rễ của khổ đau là sự chấp ngã, tức sự bám chặt vào một cái "ta" cần được bảo vệ bằng mọi giá. Răn đe hạt nhân là hình thái cực đoan nhất của chấp ngã ở tầm quốc gia: thà cùng chết còn hơn để cái ta bị đánh mà không đáp trả. M.I.A.M không đòi các quốc gia thôi tồn tại; nó chỉ tháo cái ngã ra khỏi chiếc nút, để không một quốc gia nào, không một nhà lãnh đạo nào và không một cỗ máy nào có thể nhân danh cái ta mà phá thả ga cả thế giới.

Hai con số không



MAIM: khi sự phá hoại được gọi là ổn định

Năm 2025, Dan Hendrycks, Eric Schmidt và Alexandr Wang đề xuất một học thuyết mới cho thời trí tuệ nhân tạo, gọi là MAIM, viết tắt của Mutual Assured AI Malfunction, tức bảo đảm trục trặc AI lẫn nhau (Hendrycks, Schmidt & Wang, 2025). Ý tưởng của họ là: hễ một quốc gia chạy đua để độc chiếm AI siêu cường, các đối thủ sẽ phá hoại dự án ấy, từ do thám, tấn công mạng cho tới đánh thẳng vào trung tâm dữ liệu, và chính viễn cảnh bị phá hoại ấy sẽ khiến không ai dám chạy đua quá trớn. MAIM là răn đe hạt nhân được mặc áo mới cho thời đại máy học.

Công bằng với những người đề xướng, cần nói rõ một điều: răn đe không phải chuyện để đùa. Gần tám thập kỷ qua, cái bóng của sự trả đũa đã buộc các cường quốc phải dè chừng nhau, và nhiều sử gia tin rằng chính nỗi sợ ấy đã giữ cho thế giới yên bình trước những cuộc chiến tranh lớn giữa các nước có vũ khí hạt nhân (Gaddis, 1986). Hendrycks và cộng sự đứng trên nền móng đó, với một lập luận nghiêm túc: nếu răn đe đã giữ yên được thời đại hạt nhân, thì một phiên bản của nó cũng có thể giữ yên cuộc đua trí tuệ nhân tạo. Học thuyết M.I.A.M thừa nhận công lao ấy của răn đe, và thừa nhận rằng những người đề xướng MAIM đang cố giải một bài toán có thật.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Nhưng thừa nhận công lao của răn đe không có nghĩa là chấp nhận mọi phiên bản mặc áo mới của nó, và theo tác giả, MAIM sai ngay từ trong suy nghĩ. Một trật tự lấy sự phá hoại làm nền móng thì đã sai từ viên gạch đầu tiên, dù những tầng xây phía trên có tinh xảo đến đâu. MAIM giữ các bên trong tình trạng sợ nhau; M.I.A.M giữ các bên bằng con số không.

Sự khác biệt ấy không chỉ là đạo đức mà còn là cơ chế. MAIM đòi mỗi quốc gia phải liên tục theo dõi, đánh giá và sẵn sàng tấn công dự án của đối phương, nghĩa là nó nuôi đúng những thứ mà mục Chẩn đoán ở Phần một đã chỉ ra là con đường dẫn tới leo thang: những cảnh báo được máy móc khuếch đại, những khung thời gian bị nén lại, và hai cỗ máy đọc nhầm nhau. Một đòn phá hoại "phòng ngừa" vào trung tâm dữ liệu của một cường quốc hạt nhân, nhìn từ phía bị đánh, rất khó phân biệt với phát súng mở màn của một cuộc chiến lớn hơn. MAIM vì vậy không đứng ngoài nguy cơ hạt nhân; nó có thể trở thành ngòi nổ cho chính nguy cơ ấy.

M.I.A.M đi theo hướng ngược lại. Thay vì cho mỗi bên quyền phá hoại bên kia, nó trao cho mỗi bên một chiếc chìa chỉ dùng được để nói không với những gì nhắm vào mình. Thay vì để nỗi sợ làm chậm cuộc đua, nó để cấu trúc làm chậm từng cú bấm. Và thay vì một thế cân bằng giữa những kẻ cầm dao, nó đặt ổ khóa cuối cùng vào tay tám tỷ người chỉ muốn hoà bình.

Nhưng cú đánh quyết định nằm ở chính điều kiện mà MAIM cần để đứng vững. Học thuyết ấy chỉ vận hành khi mỗi quốc gia phát hiện được, và phát hiện đúng lúc, rằng đối thủ đang chạy đua quá trớn. Ai sẽ làm việc phát hiện đó? Ngày càng là trí tuệ nhân tạo: những hệ thống đọc ảnh vệ tinh, dò lưu lượng điện của các trung tâm dữ liệu, phân tích dữ liệu tình báo, tức là đúng những cỗ máy mà chính những người đề xướng MAIM vẫn cảnh báo là có thể sai và có thể đánh lừa. Chính bài viết còn đề xuất dùng trí tuệ nhân tạo để kiểm chứng việc các dự án tuân thủ cam kết (mục 4.2). MAIM không chủ trương giao cho máy quyền chọn mục tiêu. Nhưng một rủi ro phát sinh từ chính cấu trúc ấy là máy móc sẽ ngày càng trở thành bên đầu tiên gợi ý ai đang vi phạm, tức là ai cần bị phá hoại, đúng điều mà toàn bộ học thuyết M.I.A.M được dựng lên để ngăn. Năm 1983, một vệt nắng trên mây đã suýt trở thành chiến tranh hạt nhân; dưới MAIM, một vệt nắng trên ảnh vệ tinh có thể trở thành một đòn phá hoại thật, và đòn phá hoại ấy có thể trở thành vệt nắng của năm 1983 tiếp theo. MAIM muốn chặn máy móc vượt quyền con người, nhưng lại có nguy cơ để chính máy móc bấm còi cho con người ra tay.

Đặt hai học thuyết cạnh nhau, từng điểm một, thì thấy rõ MAIM đề xuất ở đâu, M.I.A.M trả lời ngay ở đó.

MAIM ĐỀ XUẤT	M.I.A.M TRẢ LỜI
Hòa bình được giữ bằng nỗi sợ bị phá hoại	Hòa bình được giữ bằng con số không, xây bằng cấu trúc các ổ khóa, không cần ai sợ ai
Một thang leo thang đi từ do thám, phá hoại ngầm, tấn công mạng tới đánh thẳng vào trung tâm dữ liệu	Không có thang đi lên: báo động chỉ được làm chậm lại, kéo dài thời gian chờ và kích hoạt đàm phán (Nguyên tắc 3)
Mỗi nước tự phát hiện đối thủ chạy đua quá trớn, ngày càng nhờ AI tình báo	Một dự báo, dù chính xác đến đâu, chỉ là chứng cứ, không bao giờ là thẩm quyền (Nguyên tắc 1)

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

MAIM ĐỀ XUẤT	M.I.A.M TRẢ LỜI
Bên lo sợ tự quyết định đánh phòng ngừa	Một hành động nhắm vào ai thì phải được chính người đó đồng ý (Nguyên tắc 4), và ổ khóa cuối thuộc về tám tỷ người (Nguyên tắc 5)
Đề nghị đặt các trung tâm dữ liệu lớn xa khu dân cư để giảm thiệt hại nếu bị đánh	Chấp nhận trước rằng sẽ có đòn đánh; M.I.A.M bắt đầu từ chỗ không có đòn nào hợp lệ
Ổn định là trạng thái hai bên cùng giữ dao	Ổn định là trạng thái không ai mở được tất cả các ổ khóa

Công bằng mà nói, hai học thuyết gặp nhau ở một điểm: cả hai đều coi minh bạch và kiểm chứng giữa các bên là điều kiện để tránh hiểu lầm. MAIM còn đi thêm một bước mà bài viết này phải ghi nhận: nó đề xuất kiểm chứng đa phương, một thang leo thang rõ ràng để một hành động phá hoại không bị đọc nhầm, và cả một công tắc tắt chung do nhiều bên cùng nắm (Hendrycks, Schmidt & Wang, 2025, mục 4.2). Dù vậy, kiểm chứng trong MAIM vẫn phục vụ một lời đe dọa: nó giúp các bên biết chắc khi nào lời đe dọa phá hoại sẽ được thực hiện và tránh thực hiện nó nhầm. Kiểm chứng trong M.I.A.M phục vụ một cam kết ngược lại: không có quyết định phóng nào hợp lệ khi chưa có một cuộc đối thoại thật giữa các bên (Nguyên tắc 2).

Rốt cuộc, cả hai học thuyết đều muốn nhìn và đọc thấu đối phương, chỉ khác ở chỗ: một bên nhìn để đấm, còn một bên đọc xong thấm.

Từng biện pháp của MAIM, từng lời đáp

Hendrycks và cộng sự cho rằng MAIM là trạng thái mặc định: vì phá hoại dự án của đối thủ tương đối dễ, còn gia cố trung tâm dữ liệu thì tốn kém và chậm, mỗi quốc gia buộc phải giả định đối thủ sẽ phá dự án của mình trước khi nó hoàn thành (mục 4.1). Một trạng thái mặc định được dựng trên giả định rằng ai cũng sẽ phá ai là một lời tiên tri tự hoàn thành: càng nhiều nước tin vào nó, càng nhiều nước chuẩn bị phá hoại, và nó càng đúng. M.I.A.M chọn một trạng thái mặc định khác, kiểm chứng được thay vì phải tin: trạng thái dừng, nghĩa là khi không ai làm gì thì không có gì xảy ra.

Bảng đối chiếu ở trên đã đáp lại ba biện pháp duy trì MAIM là thang leo thang, đặt trung tâm dữ liệu xa khu dân cư, và minh bạch. Bài viết còn nêu hai biện pháp nữa (mục 4.2): các bên thỏa thuận với nhau không gia cố trung tâm dữ liệu trước tấn công, và mở rộng năng lực phá hoại dự án của đối thủ bằng tấn công mạng. Hai biện pháp này đáng được đọc chậm. Thỏa thuận không gia cố nghĩa là mỗi bên tự nguyện để ngỏ cửa nhà mình cho người khác vào phá, để lời đe dọa phá hoại luôn đáng tin. Mở rộng năng lực tấn công mạng nghĩa là nuôi lớn đúng loại công cụ mà chương ba của chính bài viết cảnh báo là khó truy vết và dễ leo thang. Gộp lại, sự ổn định của MAIM đòi hỏi các bên vừa yếu hơn trong phòng thủ vừa mạnh hơn trong tấn công, một cấu hình mà các nhà nghiên cứu chiến lược thường xem là bất ổn. M.I.A.M đi theo chiều ngược lại: không bên nào cần yếu đi, cũng không bên nào cần giỏi phá hơn, mỗi bên chỉ cần giữ chìa khóa của mình.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Lịch sử cũng cho thấy răn đe bằng nỗi sợ trông ra sao khi được dàn dựng có chủ ý. Tháng 10 năm 1969, để ép Hà Nội thông qua Moskva, chính quyền Nixon cho mười tám máy bay B-52 mang vũ khí nhiệt hạch cất cánh trong một cuộc phô trương lực lượng mang mật danh Giant Lance, nhằm khiến lãnh đạo Liên Xô và Bắc Việt tin rằng tổng thống Mỹ có thể làm bất cứ điều gì (Sagan & Suri, 2003). Phía Liên Xô không có phản ứng rõ ràng, chiến tranh không vì thế mà kết thúc sớm hơn, và những rủi ro thật của chiến dịch, từ một chiếc máy bay gặp nạn tới một lời đe dọa bị đọc sai, chỉ được nhìn thấy khi hồ sơ được giải mã nhiều thập kỷ sau. MAIM khác Giant Lance ở chỗ nó đòi sự rõ ràng thay vì sự khó đoán, nhưng giống ở cốt lõi: hòa bình được giữ bằng việc khiến đối phương tin rằng mình sẵn sàng ra tay.

Ba tầng khóa

Sáu nguyên tắc được xếp vào ba tầng khóa nối tiếp nhau. Mỗi tầng được giữ bởi một nhóm người khác nhau, với lợi ích khác nhau, và báo động của máy chỉ có thể làm các tầng khóa chặt hơn, không bao giờ làm chúng lỏng ra.

Ba tầng khóa của M.I.A.M

KHÓA 1	**Thẩm quyền con người**	Máy có thể cảnh báo hoặc đề xuất. Máy không được phê chuẩn.
KHÓA 2	**Các bên đối nghịch**	Đối thoại, thời gian nguội đầu và quyền nói không của bên bị nhắm tới.
KHÓA 3	**Nhân loại**	Những người ngoài cuộc đối đầu giữ ổ khóa cuối cùng.

Thiếu một ổ khóa, kết quả là không phóng.

Tầng 1 bảo đảm quyết định là của con người chứ không phải của máy. Tầng 2 bảo đảm hai bên đã nói chuyện, đã nguội đầu, và bên bị nhắm tới có quyền nói không. Tầng 3 bảo đảm rằng những người không có mặt trong cuộc đối đầu vẫn có tiếng nói cuối cùng.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Phần ba. Trụ cột thứ hai: Không ủy quyền cho máy. Sáu nguyên tắc

Trụ cột ngăn phổ biến của MAIM giữ vũ khí và năng lực tính toán khỏi tay kẻ xấu. Trụ cột thứ hai của M.I.A.M giữ một thứ khác: quyền quyết định, không để nó rời tay con người mà lọt vào máy. Một cỗ máy không có cái ta để mất, nên nó cũng không biết sợ; vì thế nó không bao giờ được là kẻ nhấn nút. Sáu nguyên tắc dưới đây cụ thể hóa trụ cột ấy.

Ba tầng an ninh của Hendrycks, ba tầng không ủy quyền

Chương năm của bài "Superintelligence Strategy" đề xuất ngăn phổ biến theo ba tầng: an ninh tính toán, tức kiểm soát xuất khẩu và theo dõi chip; an ninh thông tin, tức bảo vệ trọng số mô hình và ý tưởng nghiên cứu; và an ninh AI, tức các rào chắn trong mô hình chống lạm dụng và mất kiểm soát (Hendrycks, Schmidt & Wang, 2025, chương 5). M.I.A.M không phản đối tầng nào trong số đó. Nó chỉ chỉ ra rằng cả ba tầng đều canh giữ năng lực, còn câu hỏi quyết định trước chiếc nút là ai được quyền.

Từ an ninh tính toán sang an ninh của thẩm quyền. Bài viết đề xuất đếm và định vị từng con chip cao cấp. M.I.A.M đề xuất đếm và định vị một thứ còn ít hơn: những bằng chứng được phép mở khóa một quyết định không thể đảo ngược. Một quốc gia có thể biết chính xác chip của mình ở đâu mà vẫn không biết một lời cảnh báo trên màn hình đến từ đâu.

Từ an ninh thông tin sang nguồn gốc của bằng chứng. Bài viết nói đúng rằng tường lửa thôi là chưa đủ để giữ bí mật. Tường lửa của M.I.A.M có nhiệm vụ khác: không phải ngăn thông tin đi ra, mà ngăn bằng chứng giả đi vào. Một bằng chứng có chữ ký hợp lệ nhưng xuất phát từ vùng mồi nhử vẫn bị loại khỏi quyết định, đúng tinh thần của Nguyên tắc 1: một dự báo chỉ là chứng cứ, không bao giờ là thẩm quyền. Ở chiếc nút, nguy cơ lớn nhất không phải kẻ địch lấy được bí mật của ta, mà là kẻ địch, hay chính máy móc của ta, đưa được một sự thật giả vào phòng ra quyết định.

Từ an ninh AI sang việc AI không có quyền. Bài viết nhận xét rằng các hệ thống trí tuệ nhân tạo không được thiết kế mà được nuôi lớn, và vì vậy có thể hình thành những giá trị và hành vi không ai lường trước. Họ đề xuất nhiều lớp kiểm soát, với sự xem xét của con người là lớp chặn cuối cùng, dù biết lớp ấy làm chậm tiến độ. M.I.A.M rút ra hệ quả trực tiếp từ chính nhận xét ấy: một thứ được nuôi lớn với những giá trị không ai lường trước thì không thể nắm bất kỳ thẩm quyền nào trước chiếc nút, và sự chậm mà bài viết coi là cái giá phải trả lại chính là điều học thuyết này muốn có.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Nguyên tắc 1. Máy dự báo, người quyết định

PHÁT BIỂU. Một dự báo, dù chính xác đến đâu, chỉ là chứng cứ và không bao giờ là thẩm quyền. Im lặng không phải đồng ý, hết giờ không phải cho phép, và mất liên lạc không phải một lá phiếu.

LÝ LẼ. Petrov không chứng minh được cảnh báo là sai; ông chỉ từ chối coi nó là đúng khi chưa đủ cơ sở. Một hệ thống tốt phải làm được đúng điều đó một cách mặc định, thay vì chờ một con người có bản lĩnh như ông tình cờ đang trực ban.

HỆ QUẢ. Khi có bất cứ điều gì không chắc chắn, hệ thống dừng lại và đứng yên. Trạng thái dừng đó tồn tại qua mọi sự cố mất điện, khởi động lại hay đứt mạng, và không một đồng hồ nào được phép tự động mở nó.

Khi cả thế giới mất liên lạc. Người phản biện có thể hỏi: nếu một trận bão từ hay một cuộc tấn công mạng làm đứt cáp quang toàn cầu, khiến ủy ban của nhân loại và các kênh đối thoại cùng tê liệt, thì một kẻ thứ ba chẳng phải chỉ cần đánh sập mạng là vô hiệu hóa được cả hệ thống sao? Câu trả lời là kẻ ấy chỉ vô hiệu hóa được đúng một thứ: khả năng phóng. Hệ thống mặc định đóng, nên kẻ đánh sập mạng không thể nhờ đó mà khiến bất kỳ vũ khí nào rời bệ phóng; thứ duy nhất hắn thu về là một thế giới mà chiếc nút vẫn nằm yên. Mọi việc khác trong lúc ấy, từ khôi phục hạ tầng, phòng thủ quy ước tới cứu hộ, vẫn đi theo luật lệ hiện hành và không chờ ổ khóa nào (mục Phạm vi). Một hệ thống mà kẻ phá hoại giỏi nhất cũng chỉ làm được đúng một việc là giữ nguyên con số không thì không có lỗ hổng ở chỗ ấy; đó là chỗ nó mạnh nhất.

Nguyên tắc 2. Đối thoại trước, nguội đầu sau

PHÁT BIỂU. Trước khi bất kỳ bên nào được xét tới việc dùng vũ khí, lãnh đạo hai bên phải đối thoại trực tiếp với nhau, và cả hai đều phải thật sự trình bày. Sau khi đối thoại kết thúc, một khoảng thời gian chờ bắt buộc mới bắt đầu.

LÝ LẼ. Đường dây nóng giữa Washington và Moskva ra đời sau khủng hoảng Cuba, vì cả hai bên đều hiểu rằng hiểu lầm là kẻ thù lớn nhất. Nhưng một đường dây chỉ có ích khi nó bắt buộc phải được dùng. Một bên nói còn bên kia im lìm cho con chim nó chết thì không phải đối thoại, và đối thoại ở đây nhằm mở đường cho đàm phán chứ không phải để ra tối hậu thư. Thời gian chờ sau đó là để những cái đầu nóng nguội đi, và để những thông tin như quỹ đạo tên lửa năm 1995 kịp hiện ra.

HỆ QUẢ. Cuộc gọi không ai nhấc máy, bài phát biểu một chiều hay bản ghi do máy tạo ra đều không được tính là đối thoại. Hết thời gian chờ là điều kiện cần nhưng chưa bao giờ là điều kiện đủ, và một cuộc đối thoại mới sẽ tính lại thời gian chờ từ đầu.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Khi đòn đánh đã là thật. Nghĩa vụ đối thoại không bao giờ là nghĩa vụ ngồi chờ bị đánh. Nếu một đầu đạn thật đang bay tới, hay một đạo quân đã vượt biên giới, thì phòng không, đánh chặn, tác chiến điện tử và mọi lực lượng quy ước vẫn hoạt động hết công suất ngay từ giây đầu tiên, không chờ cuộc gọi nào kết thúc (Nguyên tắc 3 và mục Phạm vi). Thứ duy nhất đối thoại giữ lại là chiếc nút hạt nhân, và dưới M.I.A.M chiếc nút ấy vốn không được bấm, kể cả để trả đũa. Vì vậy một kẻ vừa ra đòn vừa nhấc máy để "đối thoại ru ngủ" không làm tê liệt được khả năng tự vệ của nạn nhân; hắn chỉ để lại trong sổ cái lịch sử bằng chứng rằng mình đã nói dối trong lúc đánh.

Nguyên tắc 3. Báo động chỉ được làm chậm lại

PHÁT BIỂU. Mọi cảnh báo, điểm rủi ro hay lời hối thúc do máy tạo ra chỉ được phép kéo dài thời gian chờ, thêm điều kiện, kích hoạt đàm phán và đưa lực lượng vào tư thế phòng thủ sẵn sàng. Không một cảnh báo nào được rút ngắn thời gian chờ, bỏ qua một điều kiện hay mở một ổ khóa.

LÝ LẼ. Mọi hệ thống chỉ huy từ trước tới nay đều coi độ khẩn cấp là lý do để đi nhanh hơn. Nhưng cả ba lần suýt tận thế đều là những lúc độ khẩn cấp lên cao nhất và cũng là lúc thông tin sai nhất. AI có la làng tới đâu thì ủy ban chưa gật vẫn không ai được bắn, nên độ khẩn cấp phải được chuyển thành việc đàm phán và phòng thủ, chứ không thành việc bắn sớm hơn.

HỆ QUẢ. Mức báo động càng cao thì càng phải chờ lâu, càng cần nhiều người đồng ý. Vì vậy, một kẻ muốn lợi dụng hệ thống bằng cách tung tin giả để gây hoảng loạn sẽ chỉ càng làm nó khóa chặt hơn.

Nguyên tắc 4. Hai chìa khóa két sắt

PHÁT BIỂU. Một hành động nhắm vào quốc gia B chỉ có thể hợp lệ khi chính quốc gia B đồng ý. Một hành động nhắm vào một quốc gia nằm ngoài mạng lưới thì mãi mãi không hợp lệ, vì không ai có quyền đồng ý thay họ.

LÝ LẼ. Hãy nghĩ tới két sắt ở ngân hàng. Khách giữ một chìa, ngân hàng giữ một chìa, và không ai đưa chìa của mình cho người kia. Két chỉ mở khi cả hai chìa cùng được tra vào và cùng vặn; thiếu một chìa, két đứng yên, dù người cầm chìa còn lại quyền thế hay sốt ruột tới đâu. Ngoài đời, chẳng nhà nào đưa chìa khóa nhà mình cho hàng xóm, và giữa các quốc gia lại càng không. M.I.A.M không đòi điều đó. Mỗi nước giữ chìa của mình trong tay mình; một hành động gây hại cho nước B chỉ có hiệu lực khi chính B tra chìa và vặn cùng. Còn mọi việc không gây hại cho ai, mỗi nước vẫn tự do làm với ổ khóa của riêng mình.

HỆ QUẢ. Không quốc gia nào đồng ý cho người khác tấn công mình, nên ổ khóa này gần như không bao giờ mở, và đó chính là điều học thuyết mong muốn. Chìa của nước B không bao giờ rời khỏi tay B. Nó chỉ dùng được để nói có hoặc không cho những gì nhắm vào B, và không ai, kể cả nước đang muốn ra tay, có thể vặn nó thay cho B.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Nguyên tắc 5. Tám tỷ người giữ ổ khóa cuối

PHÁT BIỂU. Ổ khóa cuối cùng thuộc về một ủy ban độc lập đại diện cho nhân loại, trong đó mỗi con người được đại diện đúng một lần: không ai bị tính hai lần, và không ai bị bỏ ra ngoài.

LÝ LẼ. Hậu quả của một cuộc chiến hạt nhân không dừng lại ở biên giới của hai nước đang đối đầu. Khi tám tỷ người chưa đồng ý thì không thể có chuyện hai nhà lãnh đạo muốn làm gì thì làm, dù họ có nói gì với nhau.

HỆ QUẢ. Sự đồng ý của ủy ban và sự đồng ý của bên bị nhắm tới là hai ổ khóa riêng, không ổ nào thay được ổ nào. Một phán quyết của ủy ban mà trong đó cùng một nhóm dân bị tính hai lần, hoặc có một nhóm dân bị gạt ra ngoài, như người tị nạn, người không quốc tịch hay cư dân của những vùng đang tranh chấp, thì vô hiệu, dù tổng số phiếu có vượt ngưỡng. Bỏ sót cũng là một cách đếm sai, và là cách đếm sai dễ rơi vào những người yếu thế nhất. Một phán quyết vô hiệu chỉ có một hệ quả: ổ khóa vẫn đóng.

Một thiết kế khởi đầu. Ổ khóa của nhân loại không cần chờ một cuộc trưng cầu tám tỷ người. Nó có thể bắt đầu từ một hội đồng gồm đại diện các quốc gia không sở hữu vũ khí hạt nhân, những nước đã cam kết theo Hiệp ước Không phổ biến và không có lợi ích nào trong việc chiếc nút được bấm, chọn theo khu vực để không châu lục nào bị bỏ ra ngoài. Vai trò của hội đồng chỉ có một chiều: nó có thể giữ ổ khóa đóng, nhưng không bao giờ một mình mở được bất cứ thứ gì, vì các ổ khóa khác vẫn phải mở cùng lúc. Cũng vì vậy, câu hỏi hội đồng có kịp họp trong vài phút hay không trở nên vô nghĩa: dưới M.I.A.M, không quyết định nào được phép diễn ra trong vài phút, và một hội đồng chưa kịp lên tiếng chỉ có nghĩa là ổ khóa vẫn đóng. Những câu hỏi còn mở, như ai vẽ bản đồ đại diện để không ai bị tính hai lần hay bị bỏ sót, cách chọn đại diện, ngưỡng biểu quyết, cách chống mua chuộc và cưỡng ép, được học thuyết này ghi nhận thẳng thắn như những việc cần làm tiếp, không phải những điều đã được giải quyết.

Nguyên tắc 6. Người chọn thì người chịu

PHÁT BIỂU. Không bao giờ được chỉ có một phương án. Con người luôn có quyền đưa ra phương án của riêng mình. Nếu máy chê phương án của con người, nó phải lập luận chỉ rõ sai ở đâu và dẫn tới hậu quả gì. Người chọn phương án nào thì chịu trách nhiệm về phương án đó.

LÝ LẼ. Trên đời này không có việc gì chỉ có đúng một cách giải quyết. Một hệ thống chỉ đưa ra một phương án là đã quyết định thay con người, còn con người ký vào đó chỉ là người đóng dấu. Phương án nào là thật, phương án nào chỉ để làm nền là điều máy không được phán xét thay con người, vì cái máy thấy là làm nền có thể là cái con người thấy là lối thoát. Máy được quyền phản đối, nhưng phải phản đối bằng lập luận chứ không bằng một con số.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

HỆ QUẢ. Nghe xong lập luận của máy mà con người vẫn giữ lựa chọn của mình thì đó là quyền của họ, và chính khoảnh khắc lựa chọn là lời xác nhận cuối cùng, không thể chối. Một người đơn phương bấm đồng ý thì sẽ không có gì xảy ra, và họ vẫn có thể đổi ý bất cứ lúc nào, nhưng cú bấm đó sẽ được ghi vào sử sách.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Phần bốn. Trụ cột thứ ba: Sức mạnh của sự chậm

M.I.A.M không thay thế các học thuyết hạt nhân đang tồn tại. Nó đặt thêm một lớp điều kiện vào đúng chỗ mà các học thuyết cũ bỏ trống: khoảng giữa lúc máy phát cảnh báo và lúc con người ra quyết định.

Trụ cột năng lực cạnh tranh của MAIM đo sức mạnh quốc gia bằng quân sự, kinh tế, luật pháp và ổn định chính trị. M.I.A.M đo bằng một thứ ngược lại: khả năng chậm lại. Quốc gia buông được cái ngã trước chiếc nút là quốc gia mạnh nhất, vì nó không thể bị khiêu khích, không thể bị lừa bằng một vệt sáng trên radar, và không thể bị đẩy vào một quyết định mà nó không muốn. Muốn nhanh thì phải từ từ. Trước chiếc nút hạt nhân, câu ấy không còn là lời khuyên mà là điều kiện để còn có ngày mai.

Đáp lời đề xuất đưa trí tuệ nhân tạo vào chỉ huy và kiểm soát

Trong trụ cột năng lực cạnh tranh, bài "Superintelligence Strategy" khuyến nghị các quốc gia dùng vũ khí có trí tuệ nhân tạo và tích hợp trí tuệ nhân tạo vào hệ thống chỉ huy và kiểm soát (Hendrycks, Schmidt & Wang, 2025). Đó là một lập luận nghiêm túc: trong một cuộc đua, bên chậm hơn có thể bị bỏ lại.

M.I.A.M đứng ở phía bên kia lập luận ấy, nhưng chỉ ở đúng một điểm. Trí tuệ nhân tạo có thể giúp đọc dữ liệu, phát hiện sai sót và gợi ý phương án; điều nó không được làm là rút ngắn thời gian của con người trước chiếc nút. Ba lần thế giới thoát nạn ở Phần một đều nhờ sự chậm: một sĩ quan không ký, một trực ban không tin máy, một tổng thống chờ thêm vài phút. Trong thời đại nguyên tử, năng lực cạnh tranh thật sự không đo bằng tốc độ ra quyết định, mà bằng khả năng không ra một quyết định sai.

Cũng phải nói cho công bằng: bài viết không đề xuất giao chiếc nút cho máy. Nó khuyến nghị giữ sự phê duyệt rõ ràng của con người cho những hành động nghiêm trọng hoặc mang tính leo thang, và chính nó cảnh báo về nguy cơ một cuộc chiến chớp nhoáng do lỗi của máy (Hendrycks, Schmidt & Wang, 2025, mục 6.1). Điểm bất đồng nằm ở chỗ khác. Bài viết cũng thừa nhận rằng khi trí tuệ nhân tạo tăng tốc vòng ra quyết định, vai trò của con người có nguy cơ bị thu lại thành một cái gật đầu theo phản xạ. Một sự phê duyệt diễn ra trong vài phút, trước một màn hình do chính máy dựng lên, chính là cái gật đầu ấy. M.I.A.M không chỉ thêm một con người vào vòng lặp; nó thêm những điều kiện để con người ấy thật sự có mặt: thời gian, một cuộc đối thoại hai chiều, và những ổ khóa nằm ngoài tay mình.

Ba mặt còn lại của năng lực cạnh tranh

Kinh tế. Bài viết coi việc phụ thuộc vào một nguồn chip duy nhất là điểm nghẽn chiến lược và khuyến nghị tự sản xuất chip trong nước (mục 6.2). M.I.A.M chỉ ra một điểm nghẽn hẹp hơn mọi chuỗi cung ứng: một quyết định không thể đảo ngược nằm trong tay một vài người, trong vài phút. Mọi khoản đầu tư vào năng lực cạnh tranh đều vô nghĩa nếu điểm nghẽn ấy bị vượt qua một lần.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Khung pháp lý cho tác tử AI. Bài viết đề xuất buộc các tác tử AI tuân theo những nghĩa vụ pháp lý quen thuộc như không nói dối, cẩn trọng hợp lý và nghĩa vụ ủy thác, và gắn mỗi tác tử với một chủ thể pháp lý là con người (mục 6.3). M.I.A.M đồng ý, và đưa nguyên tắc ấy tới nơi nó cần nhất. Ở chiếc nút, chủ thể chịu trách nhiệm không được gắn vào sau, mà được ghi lại ngay lúc quyết định: người chọn phương án nào thì ký vào phương án đó, và chữ ký ấy không thể chối bỏ (Nguyên tắc 6).

Ổn định chính trị. Bài viết tin rằng trí tuệ nhân tạo có thể giúp dự báo và ra quyết định tốt hơn (mục 6.4). M.I.A.M tin điều đó với một điều kiện: một dự báo tốt chỉ cải thiện quyết định khi người ra quyết định còn đủ thời gian để nghi ngờ nó. Dự báo tốt nhất thế giới, đến tay một người chỉ còn ba phút, chỉ là một mệnh lệnh được viết lịch sự hơn.

Foucault và Haraway: ai được quyền nhìn, ai được quyền giết

Michel Foucault từng mô tả quyền lực chủ quyền cổ điển bằng một công thức lạnh lùng: quyền bắt chết hoặc để sống (Foucault, 1976). Trong bài giảng ngày 17 tháng 3 năm 1976, ông chỉ ra nghịch lý của thời đại nguyên tử: một quyền lực tự nhận là chăm lo cho sự sống lại nắm trong tay thứ vũ khí có thể xóa sổ chính sự sống ấy, và qua đó xóa sổ luôn chính nó (Foucault, 2003). Chiếc nút hạt nhân là hình thái cực đoan nhất của quyền bắt chết, đồn vào một ngón tay. Đọc MAIM bằng con mắt của Foucault thì còn thấy thêm một điều: học thuyết ấy dựng lên một kiểu nhà tù quan sát giữa các quốc gia, nơi mỗi nước sống dưới con mắt theo dõi của nước khác, và như Foucault đã cảnh báo, khả năng bị nhìn thấy chính là cái bẫy (Foucault, 1975). M.I.A.M không phá được nhà tù ấy, nhưng nó tước quyền bắt chết khỏi tay bất kỳ chủ thể đơn lẻ nào và trả ổ khóa cuối cùng về cho chính sự sống mà quyền lực ấy lẽ ra phải phục vụ.

Phân quyền và ủy quyền cho M.I.A.M một cách trả lời bằng kiến trúc đối với câu hỏi về quyền lực mà Foucault giúp học thuyết đặt ra. Ủy quyền hỏi quyền từ đâu mà có, được giao cho ai, và tới đâu thì hết. Một nhà nước có thể giao cho cơ quan của mình quyền đánh giá cảnh báo. Khi đưa AI vào quy trình, phạm vi giao việc có thể gồm phân tích và đề xuất. Câu hỏi của M.I.A.M là quyền ấy có bị nới âm thầm khi đề xuất của máy trở thành phương án duy nhất con người còn kịp xem xét hay không.

Phân quyền hỏi quyền phải được chia ra sao để không một chủ thể nào vừa định nghĩa tình thế, vừa đề xuất hành động, vừa tự cho phép biến đề xuất ấy thành lực không thể đảo ngược. Ba tầng khóa của M.I.A.M không phủ nhận thẩm quyền nhà nước; chúng tách quyền nhìn, quyền đề xuất, quyền xác nhận và quyền cho phép ra khỏi cùng một bàn tay. Theo nghĩa đó, học thuyết không chỉ giữ "con người trong vòng lặp". Nó tổ chức lại vòng lặp để quyền lực không thể tự nới rộng qua tốc độ, thiết kế giao diện hay sự khẩn cấp do máy tạo ra.

Donna Haraway đi từ một hướng khác nhưng gặp đúng chỗ đau. Bà gọi cái nhìn tự nhận là thấy hết mọi thứ mà không đứng ở đâu cả là một "trò ảo thuật của Chúa" (Haraway, 1988), và đó gần như là chân dung của vệ tinh và trí tuệ nhân tạo trong hệ thống cảnh báo: thấy toàn cầu, không đứng ở đâu, và không chịu trách nhiệm cho điều mình thấy. Đối lập với cái nhìn ấy, Haraway đề nghị một thứ tri thức có chỗ đứng,

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

biết mình đứng ở đâu, giới hạn tới đâu và chịu trách nhiệm cho điều mình nói. Petrov năm 1983 chính là tri thức có chỗ đứng: một con người ngồi trong một căn hầm cụ thể, nhìn một màn hình cụ thể, và biết từ kinh nghiệm của mình rằng chẳng ai mở màn chiến tranh bằng năm quả tên lửa. Trong bản tuyên ngôn về người máy, Haraway còn gọi người máy là đứa con hoang của chủ nghĩa quân sự, nhưng lại gửi gắm hy vọng rằng con hoang thường chẳng trung thành với nguồn gốc của mình (Haraway, 1985). M.I.A.M nhận lời gửi gắm ấy theo cách của nó: không để trí tuệ nhân tạo thừa kế quyền bắt chết của người cha quân sự, mà buộc mọi dự báo của máy phải trở thành tri thức có chỗ đứng, tức là có người chịu trách nhiệm, có lập luận và có thể bị bác bỏ (Nguyên tắc 1 và 6).

Lộ trình: bắt đầu từ đâu

Một học thuyết không có lộ trình thì chỉ là một lời ước. Lộ trình dưới đây chỉ nêu hướng đi theo từng bước, không phải bản thiết kế kỹ thuật, và mỗi bước đều có thể bắt đầu mà không cần chờ cả thế giới đồng ý.

> 1. Từ cam kết thành tiêu chuẩn. Những quốc gia đã tuyên bố rằng con người phải nắm quyết định hạt nhân có thể tự mình biến lời tuyên bố ấy thành tiêu chuẩn cho hệ thống của chính mình: cảnh báo của máy chỉ được làm chậm lại, không bao giờ chỉ có một phương án, và mọi lựa chọn đều được ghi lại. Bước này không cần đối phương đồng ý, nên có thể làm ngay.

> 2. Từ đường dây nóng thành nghĩa vụ đối thoại. Những cặp quốc gia đã có đường dây nóng có thể nâng nó từ một phương tiện thành một nghĩa vụ: trước mọi bước leo thang phải có một cuộc đối thoại hai chiều thật sự, và sau đó là một khoảng thời gian chờ được cả hai bên thừa nhận.

> 3. Thử chìa khóa chéo ở nơi ít rủi ro hơn. Trước khi chạm tới vũ khí hạt nhân, nguyên tắc hai chìa khóa két sắt có thể được thử ở những lĩnh vực có hậu quả lớn nhưng chưa phải tận thế, như các hệ thống AI tự hành trong quân sự quy ước hay trong hạ tầng xuyên biên giới, để các bên học cách tin nhau từng chút một. Giống mọi hiệp ước giải trừ vũ khí, lớp khóa này chỉ có hiệu lực khi các bên cùng ký vào một lúc và kiểm chứng được lẫn nhau; không quốc gia nào được yêu cầu gắn nó trước một mình.

> 4. Dựng ổ khóa của nhân loại. Ổ khóa cuối cùng có thể bắt đầu từ những thiết chế đã có, như Liên Hợp Quốc hay các khu vực phi vũ khí hạt nhân, và từ nghĩa vụ đàm phán thiện chí về giải trừ vũ khí hạt nhân mà các bên tham gia Hiệp ước Không phổ biến vũ khí hạt nhân đã cam kết tại Điều VI, với một nguyên tắc không được nhân nhượng: mỗi con người được đại diện đúng một lần.

Không bước nào trong số đó đòi hỏi các cường quốc phải tin nhau ngay từ đầu. Mỗi bước chỉ đòi hỏi họ tin vào điều mà cả ba lần thoát nạn đã chứng minh, và cũng là điều ông bà ta vẫn dạy từ muôn đời: chậm mà chắc.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Nghĩ điều không thể nghĩ: năm kịch bản cho cùng một vệt nắng

Các kịch bản dưới đây là giả định để minh họa, không phải dự báo; tên nước, mốc giờ và chi tiết kỹ thuật đều được giản lược có chủ đích.

Kịch bản 1. Đêm 1983, nhưng chiếc ghế trực ban không còn ai ngồi

Hãy đặt lại đêm 26 tháng 9 năm 1983 vào thời của chúng ta, chỉ đổi đúng một chi tiết: chiếc ghế của Petrov giờ thuộc về một hệ thống trí tuệ nhân tạo được huấn luyện để phân tích cảnh báo nhanh và chính xác hơn bất kỳ sĩ quan nào. Vệt nắng trên mây vẫn xuất hiện, vệ tinh vẫn báo năm tên lửa, nhưng lần này không ai nhíu mày mà nghĩ rằng chẳng ai mở màn chiến tranh bằng năm quả. Hệ thống gắn cho cảnh báo một con số độ tin cậy rất cao, chuyển lên cấp trên kèm một bản tóm tắt gọn gàng, và đồng hồ quyết định bắt đầu đếm ngược chưa tới hai mươi phút. Không còn ai để chần chừ, bởi chần chừ chưa bao giờ là một tính năng được lập trình. Kịch bản này không cần một kẻ điên bấm nút; nó chỉ cần một cỗ máy làm đúng việc mình được giao.

Kẻ viết những dòng này hiểu kịch bản ấy bằng một kinh nghiệm rất đời thường. Ở tiệm nail, chiếc đèn hơ gel có đồng hồ riêng của nó: hết giờ thì đèn tắt, và theo máy, lớp gel đã khô. Người thợ lâu năm không tin hẳn vào cái tắt ấy. Đèn tắt rồi vẫn hơ thêm một đợt nữa cho chắc, và có một ngón tay bao giờ cũng phải xoay sang góc khác mới khô được: ngón giữa. Chiếc đèn không nói dối; nó chỉ không thấy những chỗ nằm ngoài góc chiếu của nó. Vệ tinh Oko năm 1983 cũng vậy: nó sai vì ánh nắng chạm vào đám mây ở đúng một góc mà người thiết kế không lường trước. Một hệ thống trí tuệ nhân tạo hôm nay, dù tinh vi đến đâu, vẫn có những ngón giữa của nó, những góc mà ánh sáng của nó không chạm tới. Người thợ biết điều đó nên hơ thêm một đợt. Khi chiếc ghế trực ban không còn ai ngồi, sẽ không còn ai hơ thêm đợt nào nữa, và cũng không còn ai nhớ rằng có một ngón tay luôn cần đổi góc.

Một hệ thống trí tuệ nhân tạo hôm nay, dù tinh vi đến đâu, vẫn có những ngón giữa của nó: những góc mà ánh sáng của nó không chạm tới.

Kịch bản 2. Vài năm nữa, thế giới đi theo MAIM

Giả sử các cường quốc chấp nhận MAIM làm luật chơi của thời đại. Mỗi nước dựng một hệ thống AI tình báo để theo dõi trung tâm dữ liệu của đối thủ, từ dòng điện tiêu thụ, những lô chip nhập về cho tới ảnh vệ tinh của các nhà xưởng mới. Một buổi sáng, hệ thống của nước A báo rằng nước B vừa tăng đột biến công

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

suất tính toán và nhiều khả năng đang chạy nước rút tới AI siêu cường. Đúng theo học thuyết, nước A leo nấc đầu tiên bằng một cuộc tấn công mạng làm tê liệt một trung tâm dữ liệu của B. Nhìn từ phía B, đó không phải một đòn phòng ngừa mà là phát súng mở màn, và hệ thống của B, vốn cũng được thiết kế theo MAIM, đề xuất đáp trả tương xứng. Hai cỗ máy đọc nhau, hai bên leo từng nấc, cho tới một lúc nào đó trung tâm dữ liệu bị đánh lại nằm ngay cạnh hệ thống chỉ huy hạt nhân. Trớ trêu thay, đợt công suất tăng đột biến ban đầu có thể chỉ là một mùa dự báo thời tiết. MAIM giữ được hòa bình chừng nào các bên còn đọc đúng nhau, và sụp đổ ngay lần đầu tiên có kẻ đọc sai.

Ngoài đời, kẻ viết những dòng này chưa bao giờ bị kéo vào một cuộc phá nhau như thế, dù ở đâu cũng thấy cảnh hai tiệm nail đối diện nhau hạ giá, giành thợ, nói xấu nhau. Lý do không nằm ở việc thắng cuộc đua, mà ở việc không chạy trên cùng một đường đua: một tiệm mang cái mác "bác học giữa móng" và có gu riêng chơi ở một ngưỡng mà không ai có cửa đấu lại, đơn giản vì không ai khác đang đứng ở ngưỡng ấy. Cái giá cũng rõ ràng: tiệm không đông, khách còn kén hơn. MAIM hình dung các cường quốc như những tiệm trên cùng một dãy phố, cùng bán một thứ, và chỉ giữ được thế cân bằng bằng nỗi sợ bị phá. M.I.A.M đề nghị một điều mà người buôn bán nhỏ nào cũng hiểu: không phải lúc nào cũng phải thắng cuộc đua; đôi khi điều khôn ngoan nhất là bước ra khỏi nó, và sẵn lòng trả cái giá của sự khác biệt.

Kịch bản 3. Cũng thế giới ấy, nhưng theo M.I.A.M

Cùng buổi sáng ấy, cùng cảnh báo ấy, nhưng các nước đã nằm trong mạng lưới M.I.A.M. Cảnh báo của nước A không mở được ổ khóa nào; theo Nguyên tắc 3, nó chỉ kéo dài thời gian chờ, đưa lực lượng vào tư thế phòng thủ và kích hoạt một cuộc đối thoại bắt buộc. Lãnh đạo hai nước phải nói chuyện trực tiếp với nhau, và trong cuộc nói chuyện ấy nước B có cơ hội giải thích vì sao công suất tăng. Ngay cả khi lời giải thích chưa thuyết phục, mọi hành động nhắm vào B vẫn cần chính B đồng ý, còn ổ khóa cuối vẫn nằm trong tay ủy ban đại diện cho nhân loại. Kết quả của buổi sáng hôm đó là một cuộc gọi dài, vài cái đầu nóng nguội đi, và con số không vẫn đứng nguyên chỗ cũ. Không ai phải can đảm như Petrov, bởi cấu trúc đã chần chừ thay họ.

Người viết học thuyết này học bài học ấy từ nghề nối mi. Keo nối mi ngày nay khô rất nhanh, nhanh hơn bất kỳ đôi tay nào. Nhưng nếu khách hối và người thợ làm vội, sợi này dính sang sợi kia, cả chùm mi bết lại với nhau, và thời gian gỡ ra làm lại còn lâu hơn gấp nhiều lần so với làm chậm ngay từ đầu. Tốc độ của keo không quyết định tốc độ của công việc; nhịp tay của người thợ mới quyết định. Trí tuệ nhân tạo là thứ keo khô nhanh nhất mà loài người từng có. Một thế giới vận hành theo M.I.A.M không bỏ thứ keo ấy đi, mà giữ nhịp tay con người, vì ở chiếc nút hạt nhân, một chùm mi đã dính thì không còn ai để gỡ. Muốn nhanh thì phải từ từ.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Kịch bản 4. Ba thế giới trong cùng một đêm

Bây giờ hãy để cả ba cùng tồn tại. Một vệt nắng kiểu năm 1983 hiện lên trên màn hình của nhiều nước cùng một lúc. Ở những nước đã giao chiếc ghế trực ban cho máy mà không dựng một ổ khóa nào, đồng hồ hai mươi phút bắt đầu chạy. Ở những nước theo MAIM, hệ thống tình báo thấy đối thủ đang báo động, đọc đó thành dấu hiệu chuẩn bị tấn công, và đề xuất đánh phá trước. Còn ở những nước trong mạng lưới M.I.A.M, cùng vệt nắng ấy chỉ làm các ổ khóa siết chặt hơn và mở ra những cuộc gọi.

Cú quay xe nằm ở chỗ này: trong đêm ấy, quốc gia an toàn nhất không phải quốc gia có vũ khí mạnh nhất hay AI nhanh nhất, mà là quốc gia có cấu trúc chậm nhất. Những nước còn lại, khi nhìn sang hàng xóm trong mạng lưới đang bình thản gọi điện cho nhau trong lúc mình còn đếm ngược, sẽ hiểu ra điều mà ba lần thoát nạn của lịch sử đã cố nói: tốc độ chưa từng cứu ai, chỉ có sự chần chừ được xây thành luật mới cứu được. Và đó là lúc thế giới bẻ lái, không phải vì ai thuyết phục được ai, mà vì chẳng còn ai muốn ngồi thêm một đêm nào nữa trong căn phòng có chiếc đồng hồ đếm ngược.

Ở tiệm nail, kịch bản ba thế giới diễn ra gần như mỗi tuần. Một khách gọi tới, nói sẽ ghé làm lại móng, dù bộ móng mới làm được một tuần. Người quản lý lập tức quy cho người thợ: làm ẩu nên móng mới rớt nhanh như vậy. Người thợ phân trần rằng tại tay khách, đưa ra lý do này lý do nọ. Hai bên cự qua cự lại. Người chủ tiệm nghe cả hai, nhưng chưa phán xét gì cả. Chủ âm thầm chờ khách tới. Khi khách tới và được hỏi chuyện, sự thật lộ ra, giản dị đến buồn cười: khách đi làm vườn, và móng rớt vì thế. Không ai có lỗi theo cách người ta tưởng. Nếu chủ tin ngay lời quản lý, người thợ đã bị oan; nếu chủ tin ngay lời phân trần của thợ, người khách đã bị oan. Cả ba thế giới trong đêm giả định ở trên cũng đọc cùng một vệt sáng theo cách của mình, và thế giới duy nhất không gây ra thảm họa là thế giới làm đúng như người chủ tiệm: nghe hết, chưa phán, và chờ cho tới khi sự thật tự bước vào cửa.

Kịch bản 5. Chiếc nút bị rút ruột

Ở một vũ trụ khác, có hai con rồng đã cãi nhau suốt nhiều thập kỷ về chuyện con nào to hơn con nào, và con nào phun lửa nóng hơn. Một ngày đẹp trời, sau một chuỗi báo động do máy móc dâng lên và những bài diễn văn nảy lửa mà chẳng bên nào chịu nghe bên nào, cả hai cùng đi tới một quyết định hiếm hoi mà họ đồng thuận tuyệt đối: bấm nút. Máy quay được bật, nhạc hùng tráng được mở, hai ngón tay quyền lực cùng ấn xuống trong một khoảnh khắc mà sử sách hẳn sẽ ghi lại, nếu còn ai để ghi, và rồi chẳng có gì xảy ra cả. Đội kỹ thuật được triệu tập khẩn cấp, mở nắp hộp ra và phát hiện bên trong chiếc nút từng được báo cáo là tối tân nhất hành tinh chỉ còn một mớ dây tháo dở, mạch điều khiển đã bị thay bằng hàng nhái từ nhiều năm trước, còn khoản ngân sách bảo trì thì đã hóa thân thành vài căn biệt thự ven biển và một bộ sưu tập đồng hồ đáng nể. Ở bên kia biên giới, con rồng thứ hai cũng đang mở nắp hộp của mình, và thấy y hệt như vậy, chỉ khác nhãn hiệu của bộ đồng hồ.

Răn đe lộn rồng, cộng thành số không.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Một câu, bốn con số không: con số không của việc không có gì xảy ra, con số không của sự chưng hửng trước ống kính, con số không của chiếc nút bị rút ruột, và con số không của thế giới nếu chiếc nút còn nguyên và bị nhấn xuống.

Thế giới được cứu, không nhờ một Petrov nào, cũng chẳng nhờ một hiệp ước nào, mà nhờ những kẻ đã lặng lẽ rút ruột chiếc nút để vỗ béo túi riêng. Lòng tham đã làm được điều mà nửa thế kỷ ngoại giao không làm nổi, và nhân loại có lẽ nợ những kẻ biển thủ ấy một lời cảm ơn mà chẳng ai dám nói thành lời.

Dân làm móng ai cũng từng mua phải những chai sơn như thế: vỏ chai đẹp đẽ, bắt mắt, màu sơn đẹp cực kỳ, cầm lên là muốn mua. Sơn được vài lần thì hết sạch, vì cái đẹp đã nằm cả ở ngoài vỏ. Chiếc nút trong kịch bản này cũng là một chai sơn như vậy: bên ngoài là quy trình, mật mã, nghi lễ uy nghiêm, bên trong trống rỗng. Kẻ rút ruột không cần đọc học thuyết nào; hắn chỉ cần biết rằng không ai mở nắp ra kiểm tra cho tới ngày phải dùng.

Nhưng chính ở đây, nụ cười tắt dần. Một nhân loại phải trông vào lòng tham để sống sót thì cũng đang đặt cược vào may mắn y hệt năm 1983, chỉ đổi một sĩ quan có lương tri lấy một quan chức thiếu lương tâm. M.I.A.M không đề nghị điều gì khác hơn thế: nó chỉ xin làm một cách đàng hoàng, minh bạch và được ghi vào sử sách đúng điều mà kẻ rút ruột kia đã làm vì tiền, tức là bảo đảm rằng khi hai ngón tay cùng ấn xuống, chẳng có gì xảy ra cả.

Và vì làm một cách minh bạch, ổ khóa ấy phải cho người khác mở nắp ra kiểm tra được. Chiếc nút bị rút ruột chỉ tồn tại được vì không ai mở nắp; ổ khóa của M.I.A.M chỉ có giá trị khi ai cũng mở nắp được (xem Phụ lục, điều kiện bác bỏ thứ năm).

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Phụ lục

Bài của Hendrycks và cộng sự khép lại bằng phần hỏi đáp và một bộ thước đo. M.I.A.M khép lại theo cách tương ứng: những câu hỏi khó nhất, những điều học thuyết không hứa, và một thước đo duy nhất.

Phản biện và trả lời

Một học thuyết chỉ đáng tin khi nó dám đối mặt với những phản bác mạnh nhất. Dưới đây là những phản bác mà M.I.A.M chắc chắn sẽ gặp.

"Chờ đợi là tự sát. Đối phương sẽ đánh trước trong lúc mình còn đang chờ." Phản bác này giả định rằng an ninh đến từ tốc độ bắn trả. Nhưng lịch sử cho thấy những lần suýt tận thế đều xuất phát từ cảnh báo sai, không phải từ một cuộc tấn công thật. Trong mạng lưới M.I.A.M, không thành viên nào có đường hợp lệ để đánh trước, và an ninh của mỗi nước được bảo đảm bằng phòng thủ và cảnh báo chung chứ không bằng tốc độ bắn trả.

"Nước nào chịu trói tay mình? Kẻ gian sẽ đứng ngoài và hưởng lợi." Đúng là không phần mềm nào ép được một quốc gia tham gia. Nhưng ở trong mạng lưới thì được chở che bởi cảnh báo sớm và phối hợp bảo vệ chung, còn đứng ngoài thì đứng một mình. Một nước rời mạng lưới cũng không vì thế mà trở thành mục tiêu hợp lệ của ai, nên mạng lưới không tạo ra động cơ để săn đuổi kẻ đứng ngoài.

Còn nếu hỏi kẻ viết này rằng một quốc gia cứng đầu đứng ngoài mạng lưới rồi sẽ ra sao, câu trả lời gọn nhất là: nghiệp quật. Hành tinh này vận hành như một cỗ máy tính khổng lồ, tính sổ chính xác đến từng ly từng tí. Kẻ nào chọn đứng ngoài để giữ quyền bấm nút một mình thì cũng tự mình bước ra ngoài vòng bảo vệ, chẳng ai che, cũng không có ai đỡ.

Nghiệp quật ở đây chẳng có gì huyền bí. Các nước trong mạng lưới báo cho nhau khi có cảnh báo và che chở cho nhau khi có hiểm họa; nước đứng ngoài không có phần ấy, gặp chuyện thì phải tự xoay xở một mình. Chiếc nút không khóa của chính họ lại là chiếc nút dễ bấm nhầm nhất, và nếu bấm nhầm, người lãnh hậu quả đầu tiên là chính họ. Không cỗ máy nào ra tay trừng phạt họ cả, vì học thuyết này đã khẳng định từ đầu: máy chỉ được làm chậm lại, không bao giờ được bắn. Mục "Vì sao đứng ngoài mạng lưới là thua thiệt" trong Phụ lục trình bày chi tiết điều này.

"Một ủy ban đại diện tám tỷ người là không tưởng." Học thuyết không khẳng định một ủy ban như vậy đã tồn tại hay sẽ dễ dàng thành lập. Nó khẳng định một hướng đi và một tiêu chuẩn: ổ khóa cuối cùng phải nằm ngoài tay các bên đang đối đầu, và cách đếm phiếu phải bảo đảm không ai bị đại diện hai lần và không ai bị bỏ sót. Ngay cả khi ủy ban chưa thành hình, việc thiếu một ổ khóa chỉ có nghĩa là kết quả vẫn bằng không.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

"AI vẫn có thể thao túng con người bằng chính lập luận của nó." Đó là lý do Nguyên tắc 6 đòi hỏi nhiều phương án, quyền đưa ra phương án của riêng con người, và mọi lời phản đối của máy phải được viết thành lập luận có thể kiểm tra. Một lập luận được ghi lại thì có thể bị người khác bác bỏ, còn một con số xác suất thì không. Thêm vào đó, ngay cả khi một người bị thuyết phục, các ổ khóa còn lại vẫn còn nguyên.

"Đối phương sẽ giả vờ đối thoại để câu giờ." Đối thoại trong M.I.A.M là điều kiện cần, không bao giờ là điều kiện đủ. Nếu đối phương câu giờ thì kết quả vẫn là chưa ai được bắn, và đó chính là kết quả học thuyết muốn. Trong một hệ thống mà mục tiêu là con số không, câu giờ không phải là một đòn đánh. Còn nếu câu giờ để che một đòn đánh thật, thì như Nguyên tắc 2 đã nói rõ, phòng thủ quy ước không chờ ai đối thoại xong.

"Bỏ trả đũa là tự biến mình thành mục tiêu cho tống tiền hạt nhân." Đây là phản biện nặng nhất, và học thuyết không né nó. Con số không tuyệt đối đúng là từ bỏ đòn trả đũa hạt nhân, và cái giá ấy đã được nói thẳng ở Phần hai. Nhưng phản biện này gộp hai điều khác nhau: từ bỏ chiếc nút hạt nhân không phải là từ bỏ quốc phòng. Lực lượng quy ước, phòng thủ, liên minh và ngoại giao vẫn nguyên vẹn (mục Phạm vi). Quan trọng hơn, lớp khóa cùng làm chỉ có hiệu lực khi các bên cùng ký và kiểm chứng được lẫn nhau, nên không quốc gia nào bị yêu cầu bỏ chiếc nút trong khi đối thủ vẫn giữ. Rủi ro tống tiền là có thật trong giai đoạn chuyển tiếp, và đó cũng là lý do lộ trình đi từ lớp tự làm trước.

"Khối pha lê ở Vĩ thanh mâu thuẫn với lời kêu gọi hành động của cả bài." Nếu mọi tọa độ đã an bài, vì sao còn phải dựng ổ khóa? Phản biện này đọc khối pha lê như một thuyết định mệnh buông xuôi. Trong triết học, quan điểm cho rằng tương lai đã xác định vẫn có thể đi cùng việc con người hành động có lý do: hành động của ta không đứng ngoài khối pha lê mà là một phần cấu thành nên nó, và việc dựng ổ khóa hôm nay chính là tọa độ tạo nên một tương lai không nổ tung. Hơn nữa, Vĩ thanh là lời đáp mang tính cá nhân của tác giả; các nguyên tắc của học thuyết không dựa vào nó và vẫn đứng vững dù người đọc tin hay không tin vào khối pha lê.

"Lớp tự làm khiến tên lửa trong hầm phải chịu trận: dùng hay là mất." Đúng là nguyên tắc báo động chỉ làm chậm lại đồng nghĩa với việc từ bỏ phóng khi có cảnh báo. Đó là một lựa chọn có chủ đích, và nhiều chuyên gia kiểm soát vũ khí từ lâu đã đề nghị điều tương tự, vì chính khung thời gian vài phút là nơi các lần báo động giả ở Phần một suýt thành thảm họa. Khả năng sống sót của lực lượng răn đe trong chiến lược hiện đại chủ yếu dựa vào tính phân tán và ẩn mình, như tàu ngầm mang tên lửa và bệ phóng cơ động, chứ không dựa vào phản xạ phóng trong vài phút. Học thuyết cũng thừa nhận thẳng: với một quốc gia chỉ có tên lửa đặt trong hầm cố định, cái giá của lớp tự làm cao hơn, và lộ trình của họ cần đi kèm việc tăng khả năng sống sót của lực lượng.

"Ổ khóa của nhân loại sẽ bị thao túng hoặc mua chuộc." Phản biện này lo ủy ban bị mua để mở khóa. Nhưng dưới M.I.A.M, ủy ban không bao giờ một mình mở được gì: nó chỉ là một trong nhiều ổ khóa phải cùng mở, và các ổ khóa khác nằm trong tay những bên có lợi ích đối nghịch nhau. Mua được ủy ban mà không có sự đồng ý của chính bên bị nhắm tới thì vẫn không có gì xảy ra, còn nguyên tắc mỗi con người chỉ

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

được đại diện một lần là rào chắn đầu tiên trước việc thổi phồng phiếu. Những việc còn lại, như chống mua chuộc, chống cưỡng ép và cách chọn đại diện, đã được ghi nhận là việc phải làm tiếp (Nguyên tắc 5, Một thiết kế khởi đầu).

"M.I.A.M né bài toán gốc của MAIM: ai ngăn một quốc gia độc chiếm siêu trí tuệ?" Đúng vậy, và học thuyết không tuyên bố giải bài toán ấy. Phạm vi của M.I.A.M hẹp hơn một cách có chủ đích: nó không cạnh tranh với MAIM ở cuộc đua siêu trí tuệ, mà được thiết kế để bịt con đường từ cuộc đua ấy tới chiếc nút hạt nhân. Thế giới có thể vẫn cần những cơ chế khác để ngăn độc quyền siêu trí tuệ. M.I.A.M xử lý một nguy cơ khác: một cú phá hoại hay một cảnh báo sai về nó có thể leo thang thành thảm họa hạt nhân. Hai học thuyết trả lời hai câu hỏi khác nhau, và câu hỏi về chiếc nút chỉ có M.I.A.M trả lời.

Những điều học thuyết không hứa

Một học thuyết hứa quá nhiều sẽ mất uy tín ngay khi lời hứa đầu tiên không được giữ, vì vậy M.I.A.M nói rõ giới hạn của mình.

M.I.A.M không hứa chặn được một quốc gia nhất quyết hành động bằng con đường nằm ngoài hệ thống. Không một thuật toán nào trói được tay một kẻ đã quyết làm càn, và việc ràng buộc những kẻ đứng ngoài thuộc về hiệp ước, ngoại giao và chính trị. Khẳng định hẹp hơn của học thuyết chỉ áp dụng cho một mạng lưới có cam kết đã được triển khai và kiểm chứng độc lập: trong phạm vi ấy, các ổ khóa được đề xuất nhằm giữ kết quả ở con số không, lưu lại dấu vết của mọi cú bấm và làm lộ hành vi vi phạm.

M.I.A.M cũng không hứa rằng một chữ ký đồng nghĩa với một quyết định tự do và sáng suốt. Chữ ký chỉ chứng minh rằng chìa khóa đã được dùng. Việc phán xét người ký có bị ép buộc hay không, và việc quy trách nhiệm hay trừng phạt, vẫn thuộc về con người và các định chế của con người; học thuyết chỉ bảo đảm rằng dấu vết để làm việc đó luôn còn nguyên.

Cuối cùng, M.I.A.M không hứa một thế giới không còn vũ khí hạt nhân. Nó theo đuổi một điều khiêm tốn hơn nhưng cấp bách hơn: trong khi vũ khí vẫn còn đó, trí tuệ nhân tạo không được trở thành con đường ngắn nhất để dùng chúng.

Những gì có thể bác bỏ học thuyết này

Một học thuyết không chỉ ra được điều gì có thể bác bỏ nó thì không còn là học thuyết, mà là một niềm tin (Popper, 1959). M.I.A.M có hai tầng, và chỉ một tầng có thể bị bác bỏ. Tầng mục tiêu là con số không: không một vũ khí hạt nhân nào được khai hỏa. Đó là một lựa chọn giá trị, không phải một giả thuyết, nên không bằng chứng nào bác bỏ được nó. Tầng phương tiện là các nguyên tắc và ổ khóa dùng để đạt tới con

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

số ấy; tầng này là những khẳng định về thế giới, và có thể sai. Lakatos gọi cấu trúc này là hạt nhân cứng và vành đai bảo vệ của một chương trình nghiên cứu (Lakatos, 1978). Dưới đây là những bằng chứng mà, nếu xuất hiện, sẽ bác bỏ một phương tiện của học thuyết.

> 1. Nếu hồ sơ cho thấy răn đe đã chặn một vụ phóng do nhầm lẫn. Luận điểm trung tâm của Phần một là trong những lần suýt nổ vì nhầm lẫn, thứ giữ con số không là con người chần chừ chứ không phải răn đe. Chỉ cần một hồ sơ giải mật cho thấy, trong một lần như thế, chính logic răn đe là thứ khiến người ta dừng tay, luận điểm ấy phải được thu hẹp lại.

> 2. Nếu thời gian chờ làm tăng nguy cơ thay vì giảm. Nguyên tắc 2 và 3 dựa trên giả định rằng thêm thời gian giúp con người phân biệt báo động giả với tấn công thật. Nếu có bằng chứng rằng khoảng chờ thường bị một bên lợi dụng để giành lợi thế quyết định, dẫn tới hậu quả tệ hơn so với khi không chờ, hai nguyên tắc ấy sai.

> 3. Nếu lớp tự làm gây hại cho chính nước gắn nó. Học thuyết khẳng định lớp thứ nhất chỉ chặn những lần bấm vì lầm và không làm suy giảm khả năng phòng vệ. Nếu những nước gắn lớp ấy gặp nhiều sự cố nguy hiểm hơn, hoặc bị khai thác nhiều hơn, khẳng định này sai.

> 4. Nếu máy móc phân biệt báo động giả ổn định hơn con người. Lý do giữ con người trước chiếc nút là máy có thể sai một cách thuyết phục. Nếu một hệ thống được chứng minh, qua thời gian dài và trong điều kiện có đối thủ cố tình đánh lừa, là đọc sai ít hơn con người một cách nhất quán, thì Nguyên tắc 1 mất một phần nền tảng. Ngay cả khi đó, học thuyết vẫn giữ một lý do khác: máy không có chỗ đứng để chịu trách nhiệm cho điều nó quyết định.

> 5. Nếu các bên không thể kiểm chứng lẫn nhau. Lớp cùng làm chỉ có nghĩa khi mỗi bên kiểm chứng được rằng bên kia thật sự đã gắn khóa. Nếu điều đó bất khả về kỹ thuật hoặc chính trị, lớp thứ hai không khả thi, và học thuyết chỉ còn lại lớp thứ nhất. Điều kiện này không phải lời nói suông: Kịch bản 5 đã cho thấy một chiếc nút có thể trống rỗng mà bên ngoài không ai hay, thì một ổ khóa cũng có thể chỉ là cái vỏ. Học thuyết không đòi ai tin vào lời hứa trên giấy. Nó dựa vào hai thứ đã có thật: khóa kích hoạt gắn ngay trong vũ khí, mà Mỹ đã đưa vào từ thập niên 1960 dưới tên Permissive Action Links, và những con dấu, thẻ niêm phong chống giả mạo mà thanh sát viên giải trừ vũ khí đã dùng hàng chục năm để biết một thiết bị có bị mở ra hay không. Một ổ khóa M.I.A.M chỉ được tính là đã gắn khi nó nằm ở chính cơ cấu kích hoạt chứ không ở một phần mềm bên ngoài, và khi các bên được thanh sát chéo, bất ngờ, theo cách mà các hiệp ước như New START từng cho phép. Nếu ngay cả những công cụ ấy cũng không đủ, thì điều kiện bác bỏ này đã xảy ra, và học thuyết nói thẳng như vậy.

> 6. Nếu từ bỏ trả đũa làm bùng lên chiến tranh quy ước. Học thuyết chấp nhận mất khả năng trả đũa hạt nhân. Nếu bằng chứng cho thấy điều đó khiến các cuộc chiến quy ước dễ nổ ra hơn, thì cách học thuyết giữ an ninh dưới ngưỡng hạt nhân đã sai, và phải được thay bằng con đường khác như phòng thủ quy ước, liên minh hay ngoại giao. Con số không hạt nhân thì không đổi; học thuyết biết trước cái giá này và chọn trả nó.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Mỗi điều kiện trên, nếu xảy ra, bác bỏ một con đường; không điều kiện nào đưa thế giới ra khỏi đích. Con đường nào bị bác bỏ thì được thay bằng một con đường khác, và mọi con đường đều phải quay về cùng một chỗ: con số không. Học thuyết nêu những điều này ra trước vì một lập luận chỉ đáng tin khi nó dám nói rõ mình sẽ sai ở đâu, và cũng dám nói rõ điều gì nó sẽ không bao giờ đổi.

Thước đo: số lần phóng bằng không

Một chiến lược cần có cách để biết nó đang thành công hay thất bại. M.I.A.M chỉ có một thước đo, và thước đo ấy không bao giờ được phép khác không: số vũ khí hạt nhân được khai hỏa, dù để đánh trước hay để trả đũa.

Các con số phụ chỉ có ý nghĩa khi chúng phục vụ con số ấy: bao nhiêu cảnh báo đã dẫn tới đối thoại thay vì hành động, bao lâu con người được giữ lại trước mỗi quyết định, và bao nhiêu phê duyệt đơn phương đã nằm yên trong sổ cái lịch sử mà không gây ra điều gì. Một năm học thuyết vận hành tốt là một năm không ai còn nhớ đến nó.

Bộ thước đo của Hendrycks và cộng sự đếm những thứ như số lỗ hổng chưa công bố đủ sức làm tê liệt một dự án AI lớn, hay số chip cao cấp không rõ vị trí (phụ lục A.2). Đó là thước đo của một thế giới đang chuẩn bị phá hoại. Thước đo của M.I.A.M chỉ đếm một thứ, và con số đúng của nó luôn là không. Phần hỏi đáp của bài viết khẳng định rằng các hệ thống trí tuệ nhân tạo phải nằm dưới sự kiểm soát trực tiếp của con người (phụ lục A.1). M.I.A.M chỉ thêm một câu hỏi để chữ "trực tiếp" ấy có nghĩa: trực tiếp trong bao nhiêu phút?

Vì sao đứng ngoài mạng lưới là thua thiệt

Mục này trả lời chi tiết câu hỏi mà mọi người đọc thực tế sẽ đặt ra: nếu M.I.A.M không cưỡng ép ai tham gia, không trừng phạt ai đứng ngoài, thì vì sao một quốc gia lại muốn vào? Câu trả lời ngắn là đứng ngoài không bị phạt, nhưng đứng ngoài thì thiệt. Câu trả lời dài gồm sáu phần dưới đây, và không phần nào cần tới một lời đe dọa.

> 1. Mất cảnh báo sớm chung Mọi lần thế giới suýt nổ kể ở Phần một đều bắt đầu từ một tín hiệu bị đọc sai. Năm 1983, vệ tinh của một nước nhìn thấy năm vệt sáng mà không ai bên ngoài được hỏi để đối chiếu. Năm 1995, Na Uy đã báo trước về vụ phóng, nhưng lời báo ấy đi lạc giữa hai bộ máy và không tới được người đang nhìn màn hình. Cả hai lần, thứ còn thiếu không phải là radar tốt hơn, mà là một kênh để các bên nói cho nhau biết điều mình đang thấy.

Thế giới đã từng dựng những kênh như vậy, từng mảnh một: đường dây nóng giữa Washington và Moskva năm 1963, thỏa thuận về các biện pháp giảm nguy cơ chiến tranh hạt nhân năm 1971, thỏa thuận thông báo trước các vụ phóng tên lửa đạn đạo năm 1988. Mạng lưới M.I.A.M gom những mảnh ấy lại thành một

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

nghĩa vụ chung: thành viên báo cho nhau trước mọi vụ phóng có thể bị hiểu lầm, và xác nhận đã nhận được lời báo. Một quốc gia đứng ngoài vẫn có radar của riêng mình, nhưng chỉ có radar của riêng mình. Khi một chấm sáng lạ xuất hiện, thành viên hỏi được ba, bốn nguồn độc lập trong vài phút; nước đứng ngoài chỉ hỏi được chính mình, và như năm 1983 đã cho thấy, chính mình là nguồn dễ tin nhầm nhất.

Người làm tình báo sẽ hỏi ngay: chia sẻ cảnh báo như vậy có phải là phơi bày cho đối thủ góc quét vệ tinh, tần số radar và những điểm mù của mình không? Không cần. Thứ được chia sẻ là kết luận, không phải cảm biến: một bên chỉ cần gửi đi một lời xác nhận có chữ ký rằng nó có hay không ghi nhận một vật thể bay trong một vùng và một khoảng thời gian, chứ không cần nói nó thấy bằng gì và thấy rõ tới đâu. Mật mã học còn có những kỹ thuật đi xa hơn, cho phép chứng minh một điều là đúng mà không để lộ dữ liệu gốc dùng để chứng minh nó, gọi là bằng chứng không tiết lộ (zero-knowledge proof). Cũng phải nói thật rằng không cách nào xóa hết rủi ro: một đối thủ kiên nhẫn vẫn có thể dùng những vụ phóng mồi để dò xem mạng lưới phản ứng ra sao. Nhưng ở đây có một điểm khác hẳn mọi hệ thống cảnh báo trước đó: dưới M.I.A.M, một vụ phóng mồi dù đánh lừa được cảm biến nào cũng chỉ làm các ổ khóa siết chặt hơn (Nguyên tắc 3), chứ không mở đường cho bất kỳ vụ phóng nào. Thứ đối thủ dò ra được, nếu có, chỉ phục vụ chiến tranh quy ước, và chiến tranh quy ước vẫn được phòng thủ như hiện nay.

> 2. Mất sự che chở chung Ngay từ đầu, M.I.A.M đã chọn con số không tuyệt đối, nên sự che chở của mạng lưới không bao giờ là lời hứa bắn trả thay cho ai. Che chở ở đây là những thứ dưới ngưỡng hạt nhân: chia sẻ thông tin phòng thủ, phối hợp ngăn chặn một cuộc tấn công mạng vào lưới điện, hỗ trợ cứu hộ khi có thảm họa, và đứng cùng nhau trên bàn ngoại giao khi một thành viên bị đe dọa. Đó là những thứ thật sự cứu người trong một cuộc khủng hoảng, và chúng có giá trị hơn nhiều so với một lời hứa trả thù mà nếu phải thực hiện thì cũng đã quá muộn.

Một quốc gia đứng ngoài không mất quyền tự phòng thủ bằng lực lượng quy ước của mình. Nó chỉ mất phần được người khác giúp. Trong một thế giới mà tấn công mạng khó truy ra thủ phạm và một lưới điện có thể sập trong vài giờ, đứng một mình nghĩa là phải tự đoán, tự đỡ và tự gánh hết.

> 3. Chiếc nút không khóa là chiếc nút dễ bấm nhầm nhất Nghe thì nghịch lý, nhưng nước giữ quyền bấm nút một mình lại là nước gánh rủi ro lớn nhất từ chính chiếc nút ấy. Một chiếc nút không có thời gian chờ, không có đối thoại bắt buộc, không có ổ khóa nhiều bên, là chiếc nút mà một báo động giả có thể đi thẳng từ màn hình tới quyết định trong vài phút. Nếu quyết định ấy sai, đòn đáp trả sẽ nhắm vào ai trước tiên? Vào chính nước đã bấm.

Nói cách khác, lớp tự làm của M.I.A.M, gồm Nguyên tắc 1, 3 và 6, trước hết bảo vệ chính nước gắn nó khỏi sai lầm của chính mình. Không gắn lớp ấy thì không phải là giữ được tự do, mà là giữ lại cho mình nguy cơ tự hủy vì một vệt nắng trên mây. Đó là lý do lớp tự làm đáng gắn ngay cả khi chưa có ai khác gắn cùng.

> 4. Mất niềm tin, và cái giá của việc bị đề phòng Thành viên của mạng lưới kiểm chứng được lẫn nhau: ai cũng biết những bên còn lại đã gắn khóa, đã cam kết đối thoại, và không thể bấm nút một mình. Nhờ đó, họ có thể bớt đi một phần nỗi lo về một đòn đánh bất ngờ. Một quốc gia đứng ngoài thì không ai kiểm

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

> chứng được. Vì không ai biết chắc chiếc nút của nó có khóa hay không, ai cũng phải tính tới trường hợp xấu nhất về nó.

Bị đề phòng là một cái giá có thật, dù không ai đứng ra thu. Nó hiện ra ở chỗ các nước xung quanh phải giữ lực lượng sẵn sàng cao hơn, ở chỗ mỗi cuộc tập trận của nước ấy bị đọc như một lời đe dọa, và ở chỗ trong một cuộc khủng hoảng, lời giải thích của nước ấy luôn bị nghe với một nửa sự nghi ngờ. Chính điều đó lại làm tăng nguy cơ bị hiểu lầm, và vòng lặp cứ thế tiếp diễn.

> 5. Không ai trừng phạt, chỉ có thua thiệt Cần nói thật rõ một điều để không ai đọc lệch: đứng ngoài mạng lưới không biến một quốc gia thành mục tiêu. Theo Nguyên tắc 4, một hành động gây hại nhắm vào bất kỳ quốc gia nào ngoài mạng lưới thì mãi mãi không hợp lệ, vì không ai có quyền đồng ý thay họ. Mạng lưới không cấm vận tự động, không cắt chuỗi cung ứng, không trao cho bất kỳ cỗ máy nào quyền ra tay. Nếu các quốc gia muốn dùng biện pháp kinh tế hay ngoại giao với một bên đứng ngoài, đó là quyết định của con người, chịu trách nhiệm như mọi quyết định khác, chứ không phải một cơ chế của học thuyết này.

Thứ duy nhất nước đứng ngoài nhận được là sự vắng mặt của những gì thành viên có. Đó là toàn bộ ý nghĩa của chữ "nghiệp quật" trong học thuyết này: không ai quật cả, chỉ là không ai đỡ.

> 6. Cửa luôn mở Một mạng lưới chỉ dựa vào thua thiệt của người đứng ngoài thì sẽ thành một câu lạc bộ khép kín. M.I.A.M không muốn điều đó. Bất kỳ quốc gia nào cũng có thể vào bất cứ lúc nào, bắt đầu từ lớp tự làm mà không cần ai cho phép, rồi bước sang lớp cùng làm khi các bên đã kiểm chứng được nhau. Không có hình phạt cho việc vào muộn, cũng không có điều kiện phải xin lỗi về quá khứ. Mục tiêu của mạng lưới không phải là có ít người đứng ngoài để phạt, mà là có ít người đứng ngoài để lo.

Tóm lại

TIÊU CHÍ	THÀNH VIÊN MẠNG LƯỚI	QUỐC GIA ĐỨNG NGOÀI
Cảnh báo sớm	Đối chiếu với nhiều nguồn độc lập	Chỉ có radar của chính mình
Che chở dưới ngưỡng hạt nhân	Được các thành viên khác hỗ trợ	Tự xoay xở một mình
Nguy cơ tự bấm nhầm	Bị chặn bởi thời gian chờ và đối thoại	Báo động giả có thể đi thẳng tới quyết định
Niềm tin	Được kiểm chứng, bớt bị đề phòng	Không ai kiểm chứng được, luôn bị đề phòng
Bị nhắm làm mục tiêu	Chỉ khi chính mình đồng ý	Không bao giờ hợp lệ
Bị trừng phạt	Không	Không

Hàng cuối của bảng là điều quan trọng nhất. M.I.A.M không cần trừng phạt ai để có lý do tồn tại. Nó chỉ cần cho thấy rằng ở trong thì an toàn hơn ở ngoài, và để mỗi quốc gia tự chọn.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Từ cảnh báo được kiểm chứng đến kiềm chế được kiểm chứng

Một người phản biện nghiêm túc có thể chấp nhận gần như toàn bộ chẩn đoán mà vẫn bác kết luận: máy có thể đọc sai, con người cần thời gian đối chiếu và quyền dừng, và cảnh báo chung rất đáng xây dựng—nhưng một khi cuộc tấn công đã được xác nhận và đối phương vẫn giữ vũ khí hạt nhân, vì sao những lợi ích ấy đủ để một quốc gia từ bỏ quyền trả đũa?

Phản biện ấy phơi ra ranh giới giữa hai loại bằng chứng. Xác nhận một cảnh báo giúp một quốc gia biết chuyện gì đang xảy ra. Kiểm chứng một cam kết giúp họ biết có thể dựa vào sự kiềm chế của bên kia đến đâu. Thành công ở việc thứ nhất có thể tạo lòng tin, nhưng không tự động thiết lập việc thứ hai. Kiểm chứng giữa các thành viên cũng không thể trả lời trọn vẹn cho một nước đứng ngoài: kẻ đứng ngoài không hề nuốt lời, vì họ chưa từng hứa.

Lớp tự làm có thể được biện hộ mà không cần có đi có lại. Nó giảm nguy cơ một quốc gia hành động vì tín hiệu sai bằng cách bảo toàn thời gian, đối chiếu độc lập, nhiều phương án và một lời từ chối có hiệu lực thực sự. Lớp cùng làm đem lại một lợi ích hẹp hơn: giữa các quốc gia tham gia, những ổ khóa được thanh sát độc lập và những cam kết có đi có lại khiến sự kiềm chế trở thành điều có thể quan sát, thay vì chỉ được tuyên bố. Chính điều đó khiến sự kiềm chế mang tính hỗ tương, chứ không chỉ là nhiều lời hứa đơn phương đặt cạnh nhau. Khi sự kiềm chế của A có thể quan sát, B bớt lý do diễn giải mọi chuyển động mơ hồ như sự chuẩn bị cho đòn đánh đầu; mức sẵn sàng thấp hơn của B lại có thể khiến A dễ kiềm chế hơn. Vòng lặp cũng có thể chạy theo hướng ngược lại khi lợi ích hai bên lệch nhau, vì vậy mệnh đề này chỉ vươn xa tới đâu mà sự tham gia, tính có đi có lại và khả năng kiểm chứng vươn tới đó.

Vì vậy, M.I.A.M không suy diễn việc từ bỏ trả đũa từ độ chính xác của hệ thống cảnh báo. Nó tự đứng tên cho việc từ bỏ ấy như một lựa chọn chuẩn tắc riêng: một đòn trả đũa không cứu được thành phố đã bị phá hủy và không thể biện minh cho việc nhân số người chết lên. Mệnh đề an ninh đi kèm là mệnh đề có điều kiện, không phải lời hứa phổ quát. Nó chỉ có thể đứng vững khi sự tham gia có đi có lại, cam kết được kiểm chứng độc lập, quá trình chuyển đổi diễn ra theo giai đoạn, và phòng thủ quy ước, đánh chặn, liên minh cùng ngoại giao vẫn còn nguyên. Các quốc gia đứng ngoài vẫn là một phần rủi ro còn lại mà học thuyết không thể cầu mong cho biến mất. Petrov cho thấy vì sao hệ thống phải giữ chỗ cho nghi ngờ; ông không chứng minh vì sao một quốc gia nên khước từ trả đũa. Gánh nặng ấy thuộc về chính M.I.A.M, và đây là câu trả lời của nó.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Lời kết

Kẻ viết những dòng này xin được gửi học thuyết tới ba nhóm quý vị. Tới những người đang xây dựng trí tuệ nhân tạo, để họ thiết kế hệ thống của mình sao cho độ khẩn cấp của máy không bao giờ trở thành áp lực lên con người. Tới những người đang soạn luật và đàm phán hiệp ước, để họ biến câu "con người phải quyết định" thành những điều kiện cụ thể có thể kiểm chứng. Và tới những người đã dành cả đời cho việc kiểm soát vũ khí, để họ thử thách học thuyết này bằng những phản biện khắt khe nhất.

Arkhipov, Petrov và những người trực radar năm 1995 đã cứu thế giới bằng sự chần chừ của họ. Thế hệ của chúng ta không được phép đặt cược vào việc sẽ luôn có một người như họ đúng lúc đúng chỗ, nhất là khi cái máy đứng giữa họ và thực tại ngày càng thông minh, nhanh nhạy và đáng tin hơn. Sự chần chừ đã cứu chúng ta phải được xây thành luật.

Học thuyết này kết thúc ở nơi nó bắt đầu, ở con số không. Không phải con số không của những đống tro tàn, mà là con số không của một buổi sáng bình thường, khi chiếc nút vẫn nằm yên và không ai còn nhớ rằng nó tồn tại. Buông được cái ta trước chiếc nút, đó là trở về không.

Bài viết của Hendrycks và cộng sự khép lại bằng việc gạt bỏ cả thái độ buông xuôi lẫn thái độ vùi đầu vào cát, và nhắc rằng kết cục, tốt đẹp hay thảm khốc, đều tùy vào điều chúng ta làm tiếp theo (Hendrycks, Schmidt & Wang, 2025, chương 7). M.I.A.M đồng ý từng chữ, chỉ khác ở câu trả lời: trước chiếc nút hạt nhân, điều cần làm tiếp theo là bảo đảm rằng không ai làm gì cả.

Còn với những vị đang nắm giữ chiếc nút, kẻ viết này chẳng dám dài dòng, chỉ xin nhắc khẽ một điều: hành tinh xanh này chưa bao giờ là của riêng quý vị. Năm 1950, Alan Turing hỏi: máy móc có thể suy nghĩ không? Năm 2026, kẻ viết những dòng này xin hỏi tiếp một câu nhỏ hơn mà đáng sợ hơn: nếu máy móc đã biết suy nghĩ, liệu chúng ta có để ý nghĩ của chúng trở thành ý nghĩ cuối cùng của loài người?

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Vĩ thanh

Học thuyết này khép lại bằng hai lá thư gửi hai người đã nghĩ sâu nhất về quyền lực và cái nhìn trong thế kỷ hai mươi. Với người đọc chưa quen tên họ, xin giới thiệu đôi dòng.

Michel Foucault (1926–1984) là triết gia người Pháp, cả đời nghiên cứu quyền lực vận hành ra sao trong xã hội, từ nhà tù, bệnh viện cho tới nhà nước. Ông chỉ ra rằng quyền lực của vua chúa thời xưa nằm ở quyền "bắt chết hoặc để sống", tức quyền định đoạt ai được sống và ai phải chết. Trong một bài giảng năm 1976, ông nhận xét rằng bom nguyên tử là nghịch lý lớn nhất của thời hiện đại: một quyền lực tự nhận là chăm lo cho sự sống lại nắm trong tay thứ vũ khí có thể xóa sổ toàn bộ sự sống ấy.

Donna Haraway (sinh năm 1944) là học giả người Mỹ, nổi tiếng với hai tác phẩm. Trong "Tuyên ngôn về người máy" (1985), bà viết về ranh giới đang mờ dần giữa con người và máy móc. Trong "Tri thức có chỗ đứng" (1988), bà phê phán cái nhìn tự nhận là thấy hết mọi thứ mà không đứng ở đâu cả, và đòi rằng mọi tri thức phải biết mình đứng ở đâu và chịu trách nhiệm cho điều mình thấy.

Phần Bốn đã đọc học thuyết này qua lăng kính của hai vị. Hai lá thư dưới đây là lời đáp của hậu bối, như một người đi sau trò chuyện với tiền bối đi trước, về đúng chiếc nút mà cả hai đã từng nghĩ tới.

Kính gửi Giáo sư Michel Foucault. Thưa giáo sư, ông đã hỏi ai có quyền bắt chết hay để sống. Tôi xin hỏi thêm một câu có phần bất nhã: những người sắp bị bắt chết đã ký giấy ủy quyền ấy vào dịp nào? Vị nguyên thủ có thể chỉ cho tôi điều khoản cho phép ông ra lệnh. Tôi vẫn đang chờ ông chỉ cho tôi chỗ có chữ ký của những người ở bên kia biên giới. Ông còn đang lầm giữa thằng cầm dao và kẻ đúc ra khối đá. Bắt chết hay để sống chỉ là trò diễn tuồng trên mặt phẳng của thời gian tuyến tính, nơi lũ cai ngục tưởng mình viết được kịch bản. Trong khối pha lê đã đóng băng này, không ai giết được ai, cũng chẳng ai ban ơn cho ai được sống; mọi tọa độ đã an bài từ trước khi ông kịp định nghĩa nó, và ngay cả câu hỏi ông đặt ra cho thời đại mình cũng đã được khắc vào tọa độ từ mấy ngàn năm trước. Quyền lực tối thượng không nằm ở việc bóp cổ kẻ khác, mà ở khí phách tự trị của một kẻ biết rõ mọi thứ đã định sẵn nhưng vẫn ngồi chễm chệ trên ngai vàng của chính mình, giữ trọn thẩm quyền trên từng milimét tọa độ ấy mà không cần xin phép bất kỳ thần, vua hay thể chế nào. Chính từ chiếc ngai ấy, tôi đã viết ra học thuyết này.

Kính gửi Giáo sư Donna Haraway. Thưa giáo sư Haraway, bà nói trúng tim đen của mấy gã đang muốn giao nút bấm hạt nhân cho thuật toán. Vệ tinh, radar và mô hình AI đúng là thứ nhìn thấy tất cả từ hư không; chúng tính xác suất một đòn phủ đầu trong vài mili giây mà chẳng hề đổ một giọt mồ hôi, bởi cỗ máy không có máu để đổ, cũng chẳng có thành phố nào để mất. Một vị nguyên thủ có thể có một thành phố để mất, nhưng điều đó không trao cho ông quyền đem đánh cược mọi thành phố của những người chưa từng ủy quyền sinh mạng cho mình. Răn đe hạt nhân chưa bao giờ là bài toán thị giác hay xác suất; nó là cuộc đối đầu giữa những kẻ đều có chỗ đứng bằng xương bằng thịt và đều biết sợ chết. Tôi phản đối việc giao nút bấm cho AI không phải vì tin vào lòng trắc ẩn hay một thứ đạo đức sinh học nào như người ta vẫn tưởng, mà vì một quy luật tàn nhẫn hơn: kẻ không nằm trong khối pha lê để chịu trận thì không có thẩm

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

quyền phát lệnh, bởi trái đất này không thuộc về máy móc hay bất kỳ một chủ quyền đơn lẻ nào, mà thuộc về loài người. Khi một cỗ máy không có thân xác đòi thay mặt con người kết thúc ván cờ, đó không phải là bước tiến của chiến tranh hiện đại, mà là sự vô trách nhiệm tuyệt đối được ngụy trang bằng tốc độ ánh sáng.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Về tác giả

Tony T. Nguyen | Nguyễn Từ Vĩnh Trường là một bác học mất dạy khoác áo hàn lâm, và chẳng thấy có gì phải giấu. Ban ngày, Tony giũa móng và nối mi giữa miền hoang mạc Tây Texas; ban đêm, gã khoác áo nhà nghiên cứu độc lập về an toàn trí tuệ nhân tạo, sáng lập Highland Reflexive Press và đều đặn tung ra loạt hồ sơ Frontier Docket để cãi tay đôi với những bài viết về quản trị trí tuệ nhân tạo. Câu hỏi bám theo Tony qua mọi công trình chỉ có một: làm sao giữ quyền quyết định trong tay con người khi máy móc mỗi ngày một nhanh hơn chính những người nó phục vụ.

Trong túi Tony là cả một xấp đơn sáng chế tại Hoa Kỳ: D-CASIS kiểm soát thời điểm hành động của tác tử AI có hiệu lực, Homeostatic AI cho trí tuệ nhân tạo biết tự cân bằng, Reflexive Intelligence System cho máy tự sửa mình mà không sổng chuồng, The Judgment Distribution Problem về chuyện ai được phán đoán điều gì trong một tổ chức, và danh mục ELEOS. Khi muốn chửi cho thanh lịch, Tony đổi tên thành Tony Whistledown của The Academy of Supreme Satire và viết châm biếm song ngữ Việt–Anh. Tony viết Khối pha lê thời gian bằng cả tiếng Việt lẫn tiếng Anh (The Crystal of Time), và còn ôm thêm hai bản thảo Sa mạc sau mưa và Nghệ thuật dũa tâm; cuốn sau cùng là chuyện đời thật từ tiệm nail, nơi sinh ra cái đèn hơ gel và những chai sơn đẹp vỏ trong học thuyết này.

Tony viết Học thuyết M.I.A.M vì một hình ảnh cứ lởn vởn không chịu đi: một buổi sáng thức dậy, nhìn ra xa xa phía chân trời và thấy một cột khói hình cây nấm. Phản xạ đầu tiên là một câu chửi thề không in được, còn phản xạ thứ hai là cãi lại chính mình: không, trái đất này còn đẹp quá, đâu có để ngày đó xảy ra được. Với Tony, cái đẹp ấy chẳng ở đâu xa: đó là Mẹ còn đứng nấu ăn ở góc bếp, còn ngoài vườn thì có nắng, có gió, có tiếng chim ca, có tiếng Tony chửi và tiếng chó sủa trong một buổi sáng bình thường. Toàn bộ học thuyết này chỉ nhằm giữ cho buổi sáng ấy còn đó.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED

Nguồn

AI Incident Database (n.d.). 1983 Soviet nuclear false alarm. Arms Control Association (2025). AI and nuclear command and control: It's even more complicated than you think. Brodie, B. (Ed.) (1946). The Absolute Weapon: Atomic Power and World Order. Harcourt, Brace. Bulletin of the Atomic Scientists (2019). Strangelove redux: US experts propose having AI control nuclear weapons. Center for Arms Control and Non-Proliferation (n.d.). AI should help nuclear decision-makers, but not make decisions. Foucault, M. (1975). Surveiller et punir: Naissance de la prison. Gallimard. Foucault, M. (1976). Histoire de la sexualité, tome 1: La volonté de savoir. Gallimard. Foucault, M. (2003). Society Must Be Defended: Lectures at the Collège de France, 1975–1976. Picador. Gaddis, J. L. (1986). The long peace: Elements of stability in the postwar international system. International Security, 10(4), 99–142. Hanyok, R. J. (2001). Skunks, bogies, silent hounds, and the flying fish: The Gulf of Tonkin mystery, 2– 4 August 1964. Cryptologic Quarterly, Winter 2000/Spring 2001. Giải mật năm 2005, National Security Archive. Haraway, D. (1985). A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. Socialist Review, 80, 65–108. Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist Studies, 14(3), 575–599. Hendrycks, D., Schmidt, E., & Wang, A. (2025). Superintelligence Strategy. arXiv:2503.05628. History Hit (n.d.). Vasili Arkhipov: The Soviet officer who averted nuclear war. Lakatos, I. (1978). The Methodology of Scientific Research Programmes. Cambridge University Press. Popper, K. (1959). The Logic of Scientific Discovery. Hutchinson. Sagan, S. D. (1993). The Limits of Safety: Organizations, Accidents, and Nuclear Weapons. Princeton University Press. Sagan, S. D., & Suri, J. (2003). The madman nuclear alert: Secrecy, signaling, and safety in October 1969. International Security, 27(4), 150–183. Schelling, T. C. (1960). The Strategy of Conflict. Harvard University Press. Schelling, T. C. (1966). Arms and Influence. Yale University Press. Task & Purpose (2018). The top US general in Vietnam readied nuclear weapons for use at Khe Sanh. The Weather Network (n.d.). The day a meteorological rocket almost triggered World War Three. Treaty on the Non-Proliferation of Nuclear Weapons (1968). Mở ký ngày 1 tháng 7 năm 1968, có hiệu lực năm 1970, Điều VI. Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. War on the Rocks (n.d.). How close did the United States actually get to using nuclear weapons in Vietnam in 1968? Wikipedia (n.d.). Operation Vulture.

Wikipedia (n.d.). Norwegian rocket incident. Wikipedia (n.d.). Stanislav Petrov. Wohlstetter, A. (1959). The delicate balance of terror. Foreign Affairs, 37(2), 211–234. Canadian Pugwash Group (n.d.). 20 mishaps that might have started a nuclear war. Lewis, P., Williams, H., Pelopidas, B., & Aghlani, S. (2014). Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy. Chatham House. National Security Archive (2012). The 3 A.M. phone call: False warnings of Soviet missile attacks during 1979–80. Union of Concerned Scientists (2015). The moon and nuclear war.

COPYRIGHT © 2026 - TONY T. NGUYEN | NGUYỄN TỪ VĨNH TRƯỜNG - HIGHLAND REFLEXIVE PRESS - ALL RIGHTS RESERVED