
# Claude, GPT, Palantir và chiến tranh thế giới năm 2026

## AI đã kéo cò chiến tranh như thế nào

Luật sư Kim Kyung-jin

AI Library - kimkj.com - 2026

# Mục lục

# Lời mở đầu

Trên màn hình hiện lên một cái tên.

Trước màn hình, một người đang ngồi. Trong căn phòng của một đơn vị tình báo nhìn xuống Dải Gaza, trên màn hình đặt trên bàn, tên và con số xuất hiện cùng nhau. Con số nằm ở đâu đó giữa 1 và 100. Số càng cao, cỗ máy càng cho rằng người này có khả năng là thành viên của một tổ chức vũ trang. Tên của cỗ máy là Lavender.

Sĩ quan nhìn cái tên. Nhìn con số. Rồi xác nhận một điều. Cái tên này có phải là tên của một người đàn ông hay không.

Nếu đúng là đàn ông, hồ sơ được thông qua.

Một sĩ quan tình báo trực tiếp làm việc đó đã nói với Yuval Abraham, phóng viên của hai cơ quan truyền thông Israel +972 Magazine và Local Call, rằng thời gian mất cho đến bước này là khoảng 20 giây. 20 giây. Một ngụm cà phê nguội đi còn lâu hơn thế. Đó là thời gian một lần đèn giao thông đổi màu. Trong khoảng thời gian ấy không có việc xem xét vì sao cỗ máy chọn cái tên đó, dựa trên căn cứ nào để cho con số ấy. Công việc của sĩ quan là xác nhận đó có phải là người thật hay không, và có phải là đàn ông hay không. Người ta nói rằng không cần hỏi gì hơn. Sĩ quan ấy mô tả việc mình làm như thế này. Tôi là con dấu cao su đóng dấu lên quyết định của cỗ máy.

Con dấu cao su. Một vật dùng để đóng nguyên những chữ đã được khắc sẵn lên giấy tờ. Dù không đọc thứ được đóng xuống là gì, bàn tay vẫn chuyển động. Sĩ quan ấy đã nhìn bàn tay của mình như vậy.

Sau 20 giây, tọa độ ngôi nhà của người mang cái tên đó được chuyển sang một hệ thống khác. Tên của hệ thống ấy là 'Where’s Daddy?'. Chương trình này chờ đến khi mục tiêu trở về nhà. Vào buổi tối, khi gia đình tụ họp, ngay khoảnh khắc người đó mở cửa bước vào, chuông báo vang lên. Khi ấy bom rơi xuống. Một sĩ quan tình báo đã nói như thế này. Chúng tôi không quan tâm đến việc giết họ khi họ ở trong các tòa nhà quân sự. Ném bom nhà ở dễ hơn nhiều. Hệ thống được tạo ra để tìm họ theo cách đó.

Trên danh sách do cỗ máy đánh dấu có khoảng 37.000 cái tên.

Lavender đã quét gần như toàn bộ cư dân Gaza. Hệ thống chấm điểm từng người trong hơn một triệu người, từ 1 đến 100. Người đó có thường xuyên đổi điện thoại di động không. Người đó gọi điện cho những ai. Người đó tham gia những nhóm chat nào. Gom những đặc điểm nhỏ như vậy lại, cỗ máy tính ra xác suất người này là thành viên của một tổ chức vũ trang. Điểm càng cao thì tên càng được đưa vào danh sách.

Các sĩ quan biết rằng khoảng 10 phần trăm danh sách đó không phải là thành viên tổ chức vũ trang. Họ biết vì đã chọn ngẫu nhiên từng người để kiểm tra. Có người chỉ có liên hệ lỏng lẻo với Hamas, có người không liên quan gì. Những người có cùng tên hoặc cùng biệt danh cũng bị cỗ máy đánh dấu. Nhân viên dân phòng hoặc cảnh sát cũng có khi bị đưa vào danh sách chỉ vì mẫu liên lạc giống nhau. Dù biết cứ mười người có thể có một người bị nhầm, các sĩ quan vẫn đóng dấu phê chuẩn, mỗi trường hợp khoảng 20 giây.

Những đợt ném bom đó có một giới hạn được định sẵn. Bộ phận luật quốc tế của quân đội đã quy định khi tấn công một mục tiêu cấp thấp thì có thể để bao nhiêu dân thường chết kèm theo. Có lời chứng nói con số đó là mười lăm, có lời chứng nói là hai mươi. Họ nói rằng chỉ cần mục tiêu là thành viên Hamas thì tuổi tác hay cấp bậc đều không được xét đến. Một sĩ quan tình báo nói rằng trong giai đoạn ấy, nguyên tắc tương xứng trên thực tế không còn tồn tại. Nguyên tắc tương xứng. Đó là lời hứa lâu đời của luật chiến tranh, yêu cầu đặt lợi ích quân sự thu được và sự hy sinh của dân thường lên bàn cân. Nghĩa là chiếc cân ấy đã bị cất đi trong một thời gian.

Đây là nơi cuốn sách này bắt đầu.

Việc một người sống hay chết được quyết định trong vòng 20 giây. Và trong 20 giây ấy, con người trên thực tế đã không hỏi lại việc mà cỗ máy làm. Điều gì đã bước vào chỗ không được hỏi ấy, hay đúng hơn, điều gì đã biến mất khỏi đó. Cuốn sách này lần theo chỗ đã biến mất ấy. Tôi sẽ gọi đó là một khoảng trống.

Tôi cần nói một chút về công việc của mình.

Tôi đã làm việc với pháp luật trong nhiều năm. Tôi biết việc truy trách nhiệm một người nghĩa là gì, và việc đó nặng đến mức nào. Truy trách nhiệm rốt cuộc quay về một câu hỏi. Ai đã làm. Ai đã quyết định làm việc đó, và ai gánh lấy hậu quả. Nếu trong phiên tòa không trả lời được câu hỏi này, vụ án không thể tiến thêm dù chỉ một bước.

Nhưng trong căn phòng ở Gaza ấy, việc trả lời câu hỏi này đã trở nên khó khăn.

Máy đã chọn một cái tên. Con người nhìn trong 20 giây. Bom rơi xuống. Một gia đình chết. Vậy ai đã quyết định giết người ấy? Là Lavender sao? Nó không phải con người, nên không thể truy trách nhiệm. Là sĩ quan đã đóng dấu chấp thuận sao? Ông ta nói mình chỉ tin vào máy, rằng mình chỉ là một con dấu cao su. Là người tạo ra Lavender sao? Ông ta chỉ lập danh sách, chứ không bóp cò. Là bộ chỉ huy đã ra lệnh chuyển danh sách thành vũ khí sao? Họ không nhìn từng cái tên riêng lẻ.

Ngón tay bóp cò rõ ràng có tồn tại, nhưng khi lần theo để hỏi trách nhiệm nơi ngón tay ấy, đến một điểm nào đó con đường bị cắt đứt. Ở đâu đó giữa con người và máy móc, trách nhiệm bốc hơi. Không phải nhẹ đi, mà biến mất hẳn. Ở chỗ ấy không có ai đứng đó.

Trong số các sĩ quan tình báo từng nói chuyện với Abraham, có vài người đau khổ vì việc mình đã làm. Họ là những người bị chấn động bởi cuộc tấn công của Hamas ngày 7 tháng 10 rồi mặc

lại quân phục. Rồi họ dao động khi nhận ra mình đang làm gì. Tạo ra những ngôi nhà nơi cả gia đình biến mất. Một số người trong họ cảm thấy không thể biện minh cho việc ấy, nên đã mở lời với nhà báo. Nghĩa là đã có người cảm thấy trách nhiệm. Chỉ có điều, trách nhiệm ấy không phải do thể chế truy hỏi, mà do từng cá nhân tự gánh một mình. Hệ thống không truy trách nhiệm bất cứ ai.

Khoảng trống ấy là chủ đề của cuốn sách này.

Đây là lần đầu tiên trong lịch sử trí tuệ nhân tạo đi vào chuỗi tử vong của chiến tranh.

Trong quân đội, chuỗi tử vong ấy được gọi là kill chain. Đó là một loạt bước: tìm mục tiêu, xác định vị trí, theo dõi, ngắm bắn, khai hỏa, rồi kiểm tra kết quả. Trước đây, gần như mọi mắt xích của chuỗi này đều do con người tự tay kéo. Muốn tạo ra một mục tiêu, nhiều sĩ quan tình báo phải xem xét trong nhiều ngày. Bàn tay con người chậm. Sự chậm ấy đôi khi cứu được một mạng người. Trong lúc kiểm tra thêm một lần, trong lúc do dự thêm một lần, trong lúc quả bom rơi xuống muộn hơn.

Ở Gaza, sự chậm ấy đã biến mất. Một sĩ quan cấp cao từng đứng đầu Trung tâm AI của Đơn vị 8200 viết trong sách rằng quân đội không thể biến con người thành mục tiêu đủ nhanh, rằng con người là nút nghẽn. Ông ta nói trí tuệ nhân tạo sẽ tháo nút nghẽn ấy. Lavender đã làm việc đó. Nó đổ ra hàng nghìn mục tiêu mỗi ngày. Việc từng mất nhiều ngày theo tốc độ con người đã thành vài giây theo tốc độ máy.

Tốc độ càng tăng, con người càng rời khỏi chuỗi ấy.

Quân đội Israel nói rằng không có hệ thống như vậy. Họ nói thứ họ dùng chỉ là công cụ hỗ trợ các nhà phân tích, chứ trí tuệ nhân tạo không tự phân biệt ai là thành viên của tổ chức vũ trang. Câu trả lời của quân đội là hệ thống tình báo chỉ là công cụ của nhà phân tích. Abraham đã đọc nguyên văn lời giải thích ấy cho các nguồn tin của mình nghe. Các nguồn tin trả lời rằng đó là lời nói dối. Abraham bổ sung thêm một chi tiết. Một sĩ quan từng lãnh đạo trung tâm AI của Đơn vị 8200 đã có bài giảng công khai tại Đại học Tel Aviv năm 2023, và tại đó ông ta trực tiếp nói rằng quân đội đã dùng trí tuệ nhân tạo để tìm ra những kẻ khủng bố vào năm 2021. Tài liệu trình bày còn có cả màn hình cho thấy máy chấm điểm con người. Một bên nói không có cỗ máy như vậy, bên kia nói họ đã ngồi trước cỗ máy ấy mỗi người 20 giây.

Cuốn sách này đặt cả hai lời nói ấy vào phần chính văn. Việc phán đoán bên nào gần với sự thật hơn xin tạm gác lại. Nhưng có một điều rõ ràng. Trong một thế giới mà ngay cả sự tồn tại của hệ thống như vậy cũng còn bị tranh cãi, việc hỏi ai chịu trách nhiệm về những gì hệ thống ấy đã làm còn khó hơn nữa. Không thể quy trách nhiệm cho một cỗ máy được nói là không tồn tại. Chính câu nói rằng nó không tồn tại lại phủ thêm một lớp lên chỗ trống ấy.

Gaza chỉ là điểm khởi đầu.

Trên những cánh đồng Ukraine, drone học cách ghi nhớ khuôn mặt con người. Đó là những cỗ máy được tạo ra để, ngay cả khi tín hiệu điều khiển bị cắt vì nhiễu điện tử, vẫn nhận ra mục tiêu vào khoảnh khắc cuối cùng rồi tự lao tới va chạm. Có người nói rằng 18 phút là đủ để tìm mục tiêu và đánh xuống. Năm 2026 tại Iran, xuất hiện lời chứng rằng việc quyết định giết một người chỉ mất 60 giây. Cùng năm ấy, trên bầu trời Venezuela, mục tiêu được xác định trên ảnh vệ tinh nơi trang web của một ngôi trường vẫn hoạt động bình thường. Từ chiến trường này sang chiến trường khác, bàn tay chọn mục tiêu đang chuyển từ con người sang máy móc. Trong lúc 20 giây thành 60 giây, rồi 60 giây lại thành 18 phút, chỗ trống đã biến mất ấy không được lấp đầy mà vẫn đi theo cùng.

Abraham cũng mang nỗi lo ấy. Trải qua cuộc chiến năm 2014 và cuộc chiến năm 2021, ông nhìn thấy một dòng chảy. Khi chiến tranh kết thúc, hệ thống đã được dùng ở đó sẽ được bán cho quân đội ở khắp thế giới. Ông nói cuộc chiến dựa trên AI này nguy hiểm cho nhân loại. Ông nói nó giúp con người thoát khỏi trách nhiệm. Khi máy tạo ra hồ sơ mục tiêu, trong đó có tên chỉ huy và cả thiệt hại dân sự dự kiến. Bề ngoài thì có đủ dáng vẻ tuân thủ luật quốc tế. Nhưng ông nói rằng mọi lời trong đó đều mất nghĩa.

Về những hệ thống lần đầu được dùng ở Gaza ấy, một nhà phân tích đã để lại một câu đáng sợ hơn. Thứ vận hành ở đó không phải là vũ khí đã hoàn thiện, mà là bản thử nghiệm. Một bản thử nghiệm được tinh chỉnh bằng cách giết người. Ông gọi nó là beta của cái chết, trong tiếng Anh là lethal beta. Từ beta vốn chỉ phiên bản thử nghiệm được chạy trước khi phát hành chính thức để tìm lỗi. Đó là giai đoạn trước khi một ứng dụng mới được đưa ra chính thức, một số người dùng được cho dùng trước để phát hiện vấn đề. Câu ấy có nghĩa là cuộc thử nghiệm đó đã được trả bằng mạng người. Điều câu nói ấy chỉ tới sẽ trở lại trước mắt chúng ta ở cuối cuốn sách. Bây giờ chỉ cần nhớ một điều. Bản thử nghiệm có thể được bán. Khi chiến tranh kết thúc, hệ thống ấy sẽ chuyển sang quân đội ở khắp thế giới. Chỗ trống được tinh chỉnh ở Gaza được xuất khẩu sang những bầu trời khác.

Có lẽ tôi sẽ không đưa ra được câu trả lời trong cuốn sách này.

Một thời đại đang đến, khi máy chọn mục tiêu nhanh hơn con người, chính xác hơn con người. Trong số các công ty tạo ra những cỗ máy ấy, có những nơi từng vạch ra ranh giới rằng ít nhất họ sẽ không vượt qua lằn ranh dùng công nghệ của mình để giết người. Chính công ty ấy lại bán công nghệ đó cho quân đội. Tôi vẫn chưa tìm ra cách khép lại mâu thuẫn này một cách gọn gàng. Tôi cũng sẽ không giả vờ rằng mình đã tìm ra.

Tuy vậy, tôi sẽ giữ lại đến cuối một cảnh. Chiếc bàn làm việc ở Gaza, một cái tên hiện trên màn hình, và 20 giây trôi qua trước nó.

Trong 20 giây ấy, một người đã chết. Và trong 20 giây ấy, trách nhiệm về quyết định giết người đó đã biến mất.

Ai đã bóp cò?

Cuốn sách này sẽ cầm theo câu hỏi ấy và bắt đầu từ Gaza.

# Phần 1 Tốc độ của máy móc Chương 1 Gaza: nhà máy ám sát hàng loạt và phê chuẩn trong 20 giây

## 1. Lavender: danh sách 37.000 người và sai số 10 phần trăm

Một cái tên hiện lên trên màn hình.

Bên cạnh cái tên ấy là một con số từ 1 đến 100. Điểm càng cao, nghĩa là cỗ máy càng xem người đó là thành viên lực lượng vũ trang. Sĩ quan tình báo nhìn tên và điểm số. Rồi ông ta chỉ kiểm tra một điều. Có phải nam giới không. Nếu đúng là nam giới, ông ta chuyển sang người tiếp theo. Thời gian ông ta dành cho mỗi mục tiêu khoảng 20 giây.

Phía bên kia màn hình là Gaza. Thông tin của hơn một triệu người được đưa vào trong máy, rồi đi ra dưới dạng điểm số. Thứ sĩ quan tình báo nhìn thấy là điểm số ấy. Không phải con người, mà là con số. Ông ta không biết con số ấy được tính như thế nào. Cũng không cần biết. Việc được giao cho ông ta không phải là nghi ngờ điểm số, mà là cho nó đi qua. Khi xong một người, người tiếp theo hiện lên, rồi người sau nữa hiện lên. Danh sách thì dài, còn cỗ máy không biết mệt.

Chỉ con người mới mệt.

Người đầu tiên đưa cảnh tượng này ra trước thế giới là nhà báo điều tra Israel Yuval Abraham. Tháng 4 năm 2024, ông đăng trên tạp chí cấp tiến Israel +972 Magazine và Local Call một bài phóng sự dài 8.000 từ về một hệ thống trí tuệ nhân tạo có tên Lavender. Bài viết được chia thành sáu bước, mỗi bước cho thấy một đoạn trong quy trình tự động hóa cao mà quân đội dùng để tạo ra các mục tiêu. Phóng sự ấy dựa trên lời kể của sáu sĩ quan tình báo Israel từng trực tiếp tham gia vào hệ thống Lavender. Trong cuộc phỏng vấn với hãng truyền thông Mỹ Democracy Now, Abraham giải thích ý nghĩa của 20 giây ấy như sau. "Một nguồn tin tình báo nói rằng ông ta dành 20 giây cho mỗi mục tiêu. Trước khi phê chuẩn vụ ném bom." Thứ ông ta kiểm tra không phải là tội của mục tiêu. Mà là giới tính.

Hãy thử cảm nhận 20 giây trong chốc lát. Đó là thời gian chờ đèn giao thông đổi màu. Là thời gian thang máy đi từ tầng này lên tầng khác. Là thời gian để một ngụm cà phê nguội bớt. Trong khoảng thời gian ấy, sinh tử của một người được phê duyệt. Kể cả gia đình người đó. Trong lúc kim giây của đồng hồ đi qua hai mươi nấc, cái tên trên màn hình trở thành mục tiêu, và ở nấc tiếp theo, cái tên khác lại hiện lên.

Trước hết, hãy xem Lavender là cỗ máy làm việc gì.

Theo tường thuật của +972, Lavender là một hệ thống trí tuệ nhân tạo do quân đội Israel tạo ra để nhận diện các nhân viên cấp thấp của Hamas và Islamic Jihad. Ý định thiết kế ban đầu khá rõ. Israel ước tính số thành viên Hamas vào khoảng 30.000 đến 40.000 người. Việc dùng con

người để lần lượt sàng lọc một số lượng lớn như vậy là bất khả thi. Vì thế họ quyết định giao việc đó cho máy. Theo cách nói của Abraham, Lavender đã rà soát thông tin của khoảng 90% cư dân Gaza, tức hơn 1 triệu người. Rồi hệ thống chấm điểm từng người một. Từ 1 đến 100. Điểm số ấy thể hiện, theo đánh giá của máy, khả năng người đó là thành viên của tổ chức quân sự Hamas hoặc Islamic Jihad.

Vấn đề bắt đầu sau đó.

Sau ngày 7 tháng 10 năm 2023, theo các bản tin, quân đội Israel đã đưa ra một quyết định. Họ coi hàng chục nghìn người có tên trong danh sách là những người có thể bị ném bom ngay trong nhà. Không chỉ người đó bị giết. Tất cả những người ở trong cùng tòa nhà, trẻ em, gia đình, đều bị cuốn vào. Abraham thuật lại như sau. "Một nguồn tin tình báo nói rằng anh ta cảm thấy vai trò của mình chỉ là đóng dấu cao su lên quyết định của máy." Một người giơ tay cho có. Máy quyết định, con người chỉ đóng dấu.

Máy có thể sai. Quân đội Israel cũng biết điều đó.

Theo tường thuật của Abraham, quân đội đã chọn mẫu ngẫu nhiên và kiểm tra từng trường hợp. Kết quả là họ biết khoảng 10% những người bị máy đánh dấu để giết không phải là chiến binh Hamas. Có người chỉ liên hệ lỏng lẻo với Hamas. Có người không liên quan gì cả. Một nguồn tin tình báo giải thích lỗi ấy phát sinh ra sao. Máy kéo vào những người có cùng tên và biệt danh với thành viên Hamas. Những người có mô thức liên lạc tương tự cũng vậy. Đó có thể là nhân viên dân phòng ở Gaza, cũng có thể là cảnh sát. Dù vậy, máy vẫn chấm điểm, và con người vẫn đóng dấu.

Ta cần cân nhắc phải hiểu tỷ lệ lỗi 10% này như thế nào. Ở một nơi yên bình, mức lỗi ấy có thể còn chịu được trong đời sống thường ngày. Bộ lọc email tự động phân loại sai một trong mười thư, ta có thể lục thùng rác và khôi phục lại. Thuật toán gợi ý sai một trong mười bộ phim, ta chỉ cần không xem bộ phim đó. Vì lỗi còn có thể sửa. Nhưng ném bom thì không thể sửa. Sau khi ngôi nhà đã sập và gia đình đã chết, việc phát hiện người đó không phải chiến binh không còn là một sự đính chính nữa. Vẫn là 10%, nhưng nếu hậu quả là cái chết, sức nặng của con số ấy hoàn toàn khác. Abraham cho biết quân đội biết về 10% đó, và dù biết, họ vẫn để hệ thống vận hành trong nhiều tuần mà không có sự giám sát có ý nghĩa.

Hãy nhìn lại bằng con số.

Số người có tên trong danh sách là khoảng 37.000 người.

Trong số đó, khoảng 10% không phải là lực lượng vũ trang.

Nếu quy đổi 10% ấy thành con người, đó là hàng nghìn người.

Con số 10% nghe gọn gàng. Thống kê bao giờ cũng gọn gàng. Nhưng bên trong nó là những người bị đưa vào danh sách chỉ vì trùng tên, hoặc vì hồ sơ liên lạc có vẻ giống nhau. Khi bom rơi

xuống mái nhà nơi họ đang ngủ, cỗ máy không thừa nhận sai sót. Cỗ máy chỉ xuất ra điểm số.

Một trong những lời chứng mà Abraham nghe được nói về cách danh tính của mục tiêu được xác nhận. Công việc nối người được Lavender chấm điểm với ngôi nhà của người đó, rồi kiểm tra xem người đó đã vào nhà hay chưa, đều đã được tự động hóa. Ông cho biết phần lớn thủ tục để sĩ quan tình báo trực tiếp quan sát ngôi nhà và xác nhận bằng mắt rằng mục tiêu có ở bên trong đã bị bỏ qua. Khi hệ thống phát cảnh báo, chính cảnh báo ấy thay cho việc xác nhận. Nếu cỗ máy nói là đúng, thì nó được coi là đúng. Mắt người lùi lại phía sau tín hiệu của máy.

Quân đội Israel đã bác bỏ bản tin này.

Lập trường của quân đội Israel được đăng trên các cơ quan truyền thông Mỹ như Washington Times là Lavender không phải là hệ thống tạo ra mục tiêu, mà là một 'cơ sở dữ liệu'. Câu trả lời chính thức của quân đội Israel mà Abraham đọc trong cuộc phỏng vấn với Democracy Now còn mạnh hơn. "Trái với các cáo buộc, quân đội Israel không dùng hệ thống trí tuệ nhân tạo để nhận diện phần tử khủng bố hay dự đoán một người nào đó có phải là khủng bố hay không. Các hệ thống thông tin chỉ là công cụ dành cho nhà phân tích."

Abraham nói rằng ông đã đọc nguyên văn câu trả lời này cho các nguồn tin của mình nghe. Phản ứng của họ rất ngắn. Đó là lời nói dối. Trong cuộc phỏng vấn, ông nói như sau. "Thông thường họ không nói dối trắng trợn đến mức này, nên tôi đã ngạc nhiên." Căn cứ mà ông đưa ra đến từ bên trong quân đội Israel. Một quan chức quân sự cấp cao lãnh đạo Trung tâm Trí tuệ nhân tạo của Unit 8200 đã có bài giảng công khai tại Đại học Tel Aviv năm 2023, và tại đó trực tiếp nói rằng quân đội Israel đã dùng một hệ thống trí tuệ nhân tạo để tìm khủng bố vào năm 2021. Abraham cho biết trong các slide của bài giảng có màn hình cho thấy hệ thống chấm điểm con người.

Cùng một cỗ máy, một bên gọi là danh sách tiêu diệt, bên kia gọi là công cụ phân tích. Cuốn sách này sẽ không trì hoãn phán đoán giữa hai cách gọi ấy. Điều cần ghi lại lúc này là sự thật rằng máy móc đã bước vào đoạn đường dẫn tới quyết định giết người, và trước câu hỏi ai đã đưa ra quyết định ấy, hai phía đang đưa ra những câu trả lời khác nhau. Vì con người đóng dấu nên đó là quyết định của con người chăng. Vì cỗ máy lập danh sách nên đó là quyết định của cỗ máy chăng. Trách nhiệm trở nên mờ nhạt ở đâu đó giữa hai điểm ấy. Chính vùng mờ đó là khoảng trống mà cuốn sách này lần theo.

Cách máy móc chấm điểm con người cũng đáng để nhìn kỹ.

Abraham cho biết Lavender chấm điểm theo một danh sách những đặc điểm nhỏ. Những tín hiệu như thói quen liên lạc của một người, việc người đó có thường xuyên đổi điện thoại hay không, người đó thuộc những nhóm trò chuyện nào, đều được đưa vào điểm số. Những tín hiệu ấy có thể giống với mẫu hành vi của lực lượng vũ trang. Nhưng cùng lúc, chúng cũng có thể giống với đời sống hằng ngày của người bình thường. Người thường xuyên đổi điện thoại

có thể là người nghèo. Người ở cùng một nhóm trò chuyện có thể chỉ là hàng xóm. Máy móc không biết sự khác nhau ấy. Máy móc nhìn thấy mẫu hình, rồi gắn điểm cho mẫu hình đó. Và điểm số ấy đã chia đôi sự sống và cái chết của một con người.

Abraham cũng cho biết kết quả của thống kê ấy đã tạo ra một cảnh tượng như thế nào. Theo số liệu Liên Hợp Quốc mà ông trích dẫn, trong hơn 6.000 người thiệt mạng ở giai đoạn đầu của cuộc chiến, hơn một nửa đến từ một số lượng gia tộc tương đối nhỏ. Điều đó có nghĩa là có những gia đình đã biến mất trọn vẹn. Một vụ ném bom nhắm vào một mục tiêu đã xóa luôn cha mẹ, anh chị em và con cái của người đó. Abraham nói đó là biểu hiện của việc đơn vị gia đình bị phá hủy. Trên danh sách, cái tên chỉ là một dòng. Ngoài hiện trường, đó là cả một gia đình.

Hãy quay lại con số 20 giây.

Trong 20 giây, con người có thể làm được gì. Đọc tên, nhìn điểm số, xác nhận giới tính. Chỉ vậy thôi. Không có thời gian để xem vì sao máy móc chọn người đó. Cũng không có thời gian để nhìn lại mẫu liên lạc, hay nghi ngờ khả năng đó là một người khác có cùng tên. Theo bài điều tra của Abraham, các sĩ quan tình báo được chỉ thị rằng chỉ cần xác nhận mục tiêu là nam giới thì phải chấp nhận nguyên khuyến nghị của Lavender. Mà không xem xét kỹ vì sao máy móc đưa ra phán đoán ấy. 20 giây không phải là thời gian thẩm tra. Đó là thời gian phê chuẩn.

Điều cuốn sách này muốn hỏi cần được nói rõ ở đây. Câu hỏi không phải là trí tuệ nhân tạo có giết người hay không. Máy bay thả bom do con người điều khiển, nút phê chuẩn cũng do con người bấm. Câu hỏi nằm ở chỗ khác. Có thể nói con người đã đưa ra quyết định hay không. Nếu điểm số do máy móc đưa ra, danh sách do máy móc lập, thời điểm do máy móc ấn định, còn con người chỉ đóng dấu trong 20 giây, thì ai là chủ nhân của quyết định ấy. Không có câu trả lời cho câu hỏi này, người phải trả lời đã biến mất, đó chính là khoảng trống.

## 2. Gospel và 'Bố ở đâu?': ngôi nhà trở thành mục tiêu

Nếu Lavender chọn con người, thì trước đó đã có một cỗ máy chọn tòa nhà.

Tên của nó là Gospel. Trong tiếng Hebrew là Habsora, nghĩa là phúc âm. Theo Plus972 và đài phát thanh công cộng Mỹ NPR đưa tin vào tháng 12 năm 2023, Gospel là một hệ thống do Đơn vị 8200 của quân đội Israel tạo ra, tự động sinh ra mục tiêu là các tòa nhà và công trình. Đó là một cỗ máy nhắm vào địa điểm, không phải con người.

Có một phép so sánh cho thấy cỗ máy này nhanh đến mức nào.

Cựu Tổng tham mưu trưởng quân đội Israel Aviv Kochavi nói rằng một nhà phân tích con người tạo ra 50 mục tiêu trong một năm.

Gospel, ông nói, tạo ra 100 mục tiêu trong một ngày.

Một năm và một ngày. 50 mục tiêu và 100 mục tiêu. Chênh lệch tốc độ sản xuất giữa con người và máy móc nằm trong hai dòng ấy. Theo bài của Plus972, trong khoảng 1.500 mục tiêu

mà quân đội Israel xác định trong cuộc xung đột Gaza năm 2021, khoảng 200 mục tiêu do Gospel chọn ra.

Một cựu nhân sự của quân đội Israel gọi hệ thống này bằng một cụm từ: nhà máy ám sát hàng loạt. Tiếng Anh là mass assassination factory. Từ then chốt là nhà máy. Nhà máy được tạo ra để phục vụ quy mô. Nhanh, nhiều, không dừng lại. Mục tiêu cũng được in ra theo cách đó. Đó chính là cụm từ trong nhan đề bài viết về Gospel mà Abraham viết trước loạt bài về Lavender.

Hãy giữ lại hình ảnh nhà máy trong chốc lát. Nhà máy làm giảm phán đoán của con người để tăng sản lượng. Việc từng do một người làm từ đầu đến cuối, nếu được chia cho nhiều máy móc đảm nhiệm, tốc độ sẽ tăng lên. Đổi lại, người nhìn toàn bộ biến mất. Công nhân đứng trước băng chuyền chỉ thấy phần đang đi qua trước mặt mình. Nhà phân tích trước Gospel cũng như vậy. Anh ta xử lý một mục tiêu trôi đến trước mình, rồi chuyển sang mục tiêu tiếp theo. Mục tiêu ấy đến từ mắt xích nào trong chuỗi nào, ở cuối chuỗi ai sẽ chết, tất cả đều nằm ngoài màn hình của anh ta. Đây là nơi đầu tiên trách nhiệm trở nên mờ đi. Dù không ai nói dối, nếu không có người nhìn toàn bộ, cũng sẽ không có người trả lời cho toàn bộ.

Nếu Gospel đánh dấu tòa nhà và Lavender đánh dấu con người, thì còn có một cỗ máy thứ ba quyết định người đó sẽ chết khi nào.

Cái tên nghe rợn người. 'Bố đang ở đâu'. Tiếng Anh là Where's Daddy.

Hãy nghĩ một lát về việc đặt cho một cỗ máy cái tên như vậy. Tên của các hệ thống quân sự thường mang vẻ đe dọa hoặc trung tính. Kiểu như Vòm Sắt, Mũi Tên, hay Chiếc ná của David. Nhưng hệ thống dùng để tìm ra mục tiêu khi người đó đã vào nhà gia đình mình lại được gắn với câu hỏi của một đứa trẻ. Những người đặt tên biết cỗ máy ấy nhắm vào điều gì. Ngôi nhà, ban đêm, gia đình. Cái tên không che giấu điều đó. Trái lại, sự nhẹ bẫng như một trò đùa trong cái tên ấy cách cái chết mà cỗ máy xử lý quá xa, đến mức nó cho thấy ở đâu đó trong dây chuyền này, cảm giác con người đã bị tê liệt.

Người dẫn chương trình Democracy Now, Amy Goodman, đã tóm tắt các bản tin về việc hệ thống này đã làm gì như sau. "Một hệ thống trí tuệ nhân tạo thứ hai có tên Where's Daddy đã theo dõi những người đàn ông Palestine có tên trong danh sách sát hại. Hệ thống này được cố ý thiết kế để Israel có thể nhắm vào họ khi mục tiêu ở nhà cùng gia đình vào ban đêm." Cố ý. Lời của nguồn tin mà Goodman trích dẫn làm rõ điểm này. "Chúng tôi không quan tâm đến việc chỉ giết họ khi họ đang ở trong tòa nhà quân sự hoặc tham gia hoạt động quân sự. Ngược lại, quân đội Israel đã ném bom họ tại nhà mà không do dự. Như lựa chọn đầu tiên. Bởi vì ném bom nhà của gia đình dễ hơn nhiều. Hệ thống được tạo ra để tìm họ trong những tình huống như vậy."

Abraham giải thích cách cỗ máy liên kết con người với ngôi nhà bằng nguyên lý vận hành của hệ thống giám sát.

Trong hệ thống giám sát đại trà có một khái niệm gọi là 'liên kết'. Khi nhập danh tính của một người, máy tính nhanh chóng nối danh tính đó với các thông tin khác. Ở Gaza, gần như ai cũng có nhà. Vì vậy hệ thống được thiết kế để tự động liên kết cá nhân với ngôi nhà. Theo giải thích của Abraham, phần lớn những ngôi nhà được liên kết với những người mà Lavender đánh dấu là thành viên vũ trang cấp thấp không phải là nơi diễn ra hoạt động quân sự. Đó chỉ là nhà có gia đình sinh sống. Where's Daddy bắt lấy khoảnh khắc người đó bước vào nhà và gửi cảnh báo cho sĩ quan tình báo. Thời điểm mục tiêu trở về bên gia đình trở thành thời điểm ném bom.

Danh sách của Lavender được chuyển sang Where's Daddy. Where's Daddy chờ họ bước vào nhà. Rồi ngôi nhà ấy bị ném bom. Abraham mô tả mối liên kết này một cách bình thản. Ngôi nhà gia đình nơi không có chiến dịch quân sự nào diễn ra, khoảnh khắc mục tiêu bước vào nhà, cảnh báo, rồi ném bom.

Chồng lên đó còn có một lựa chọn nữa. Dùng loại bom nào.

Abraham dẫn lại con số mà CNN của Mỹ đưa tin vào tháng 12 năm 2023. Theo đánh giá tình báo của Mỹ, 45 phần trăm số bom Israel thả xuống Gaza là loại bom không có thiết bị dẫn đường, thường được gọi là bom ngu. Khác với bom dẫn đường chính xác, bom không dẫn đường làm sập cả tòa nhà. Thiệt hại cho dân thường lớn hơn nhiều. Nhưng các nguồn tin nói rằng với các thành viên cấp thấp, họ chỉ dùng loại bom không dẫn đường này. Nghĩa là đánh sập nguyên căn nhà cùng với tất cả những người bên trong.

Hãy tạm làm rõ khác biệt giữa bom có dẫn đường và bom không dẫn đường. Bom có dẫn đường có thể đổi hướng trong lúc rơi để đánh trúng chính xác một điểm. Nó có thể nhắm vào một tầng, một căn phòng trong tòa nhà. Bom không dẫn đường thì cứ rơi xuống. Không thể điều khiển thật chính xác nó sẽ chạm vào đâu. Vì vậy nó phá rộng, phá sâu. Nó làm sập cả một tòa nhà. Dùng bom không dẫn đường để nhắm vào một mục tiêu là một người, gần như có nghĩa là đã chọn giết luôn tất cả những người ở quanh người đó. Nếu đã có loại bom có thể đánh chính xác mà vẫn không dùng, thì đó không phải là vấn đề năng lực. Đó là vấn đề lựa chọn. Theo các bản tin, lý do của lựa chọn ấy là chi phí.

Lý do ấy là một kiểu kế toán khô khốc đến tàn nhẫn.

Abraham kể rằng khi hỏi các sĩ quan tình báo vì sao, câu trả lời nhận được là như thế này. Những người đó không quan trọng. Họ không quan trọng đến mức đáng lãng phí loại bom đắt tiền về mặt quân sự. Nếu dùng bom có dẫn đường đắt tiền, quân đội có thể chỉ nhắm vào một tầng cụ thể của tòa nhà. Nghĩa là có thể chỉ làm sập tầng có mục tiêu, còn các gia đình khác ở tầng trên và tầng dưới vẫn có thể sống. Nhưng với các thành viên cấp thấp, họ không chi khoản tiền đó. Thay vào đó, họ dùng loại bom làm sập toàn bộ tòa nhà. Cùng với mọi tầng, mọi gia đình bên trong.

Abraham nói cảnh này đã để lại trong ông một ấn tượng sâu. Xin chuyển lại lời ông. Thả bom xuống một căn nhà và giết cả gia đình, trong khi mục tiêu mà vụ ném bom ấy nhằm ám sát lại bị coi là không đủ quan trọng để lãng phí một quả bom đắt tiền. Ông nhìn nhận đây là một cảnh hiếm hoi cho thấy quân đội Israel đo giá trị mạng sống của người Palestine như thế nào. Mục tiêu thì tầm thường đến mức chỉ đáng dùng bom rẻ, còn gia đình chết cùng mục tiêu ấy lại còn tầm thường hơn, đến mức có thể bị chôn vùi trọn vẹn bằng quả bom rẻ đó. Trong logic kế toán, sức nặng của con người được định giá như vậy.

Abraham đã chỉ thẳng mâu thuẫn ở điểm này. Xin chuyển nguyên ý lời ông. "Nếu người đó không quan trọng đến mức đáng lãng phí đạn dược, thì làm sao các ông lại sẵn sàng giết 15 thường dân?" Một mục tiêu là một người thì tầm thường đến mức không đáng dùng bom đắt tiền, nhưng để giết một người ấy, họ vẫn thả quả bom vùi chôn cả gia đình. Mâu thuẫn này vẫn nằm nguyên trong bản tin, không hề được che lấp.

Abraham cho rằng điều này cho thấy cách họ đối xử với nguyên tắc tương xứng. Tương xứng. Đây là khái niệm cốt lõi của luật nhân đạo quốc tế. Nguyên tắc ấy nói rằng thiệt hại dân sự không được quá mức so với lợi ích quân sự dự kiến. Nhưng ở đây, giá trị quân sự của mục tiêu lại được quy đổi thành giá tiền của quả bom. Rồi với mỗi mục tiêu như vậy, họ cho phép giết 15 thường dân. Theo cách nói của Abraham, đó là một cảnh hiếm hoi cho thấy quân đội Israel đo giá trị mạng sống của người Palestine ra sao khi đặt cạnh lợi ích quân sự dự kiến.

Hãy thử nhắc lại cái tên Where's Daddy.

Bố đâu rồi. Câu này thường là lời của một đứa trẻ. Đứa trẻ chờ cánh cửa mở vào buổi tối, rồi hỏi người cha vừa trở về. Nhưng cỗ máy mang cái tên ấy lại vận hành theo hướng ngược hẳn. Nó bắt lấy đúng khoảnh khắc người cha trở về nhà và biến khoảnh khắc đó thành tín hiệu ném bom. Thời điểm gia đình tụ họp trở thành thời điểm cả gia đình cùng chết. Chúng ta không biết người đặt tên ấy đã nghĩ gì. Chỉ có cảnh mà cái tên ấy chỉ tới là rõ ràng. Nhà. Buổi tối. Người vừa trở về. Và quả bom không dẫn đường rơi xuống bên trên. Theo các bản tin, trong số những người chết như vậy có nhiều phụ nữ và trẻ em.

Ông kể rằng ngay cả các nguồn tin tình báo cũng bị chấn động. Về mặt tâm lý, đó là một cú sốc. Những người trực tiếp làm việc này đã nói như vậy. Abraham cũng kể vì sao họ mở lời. Một số trong họ là những người bị thảm kịch ngày 7 tháng 10 làm cho chấn động, rồi lại được gọi nhập ngũ. Có người chưa từng nghĩ mình sẽ mặc quân phục trở lại. Rồi dần dần, họ nhận ra mình đang tham gia vào việc gì. Cảm giác rằng mình đang dính vào việc giết hại các gia đình, vào một việc không thể biện minh. Abraham kể rằng họ cảm thấy có trách nhiệm, và họ lên tiếng vì nghĩ rằng thế giới không biết sự thật. Bởi lời giải thích của người phát ngôn quân đội quá khác với thực tế tại hiện trường.

## 3. Con người trở thành người bấm nút: 20 giây xác nhận giới tính và Fire Factory

Có một cách nói rằng con người vẫn nằm trong vòng quyết định.

Tiếng Anh gọi là human in the loop. Nghĩa là con người vẫn ở lại đến cuối cùng, tại vị trí kéo cò. Đó là lời hứa rằng dù máy móc nhanh đến đâu, con người vẫn xem xét và con người vẫn phê chuẩn. Ở Gaza, con người ấy có mặt tại vị trí đó. Chỉ có điều việc anh ta làm không phải là xem xét.

Có một cuốn sách do người từng lãnh đạo Trung tâm trí tuệ nhân tạo của Đơn vị 8200 viết vào năm 2021. Tựa sách là 'The Human-Machine Team'. Sách có phụ đề nói về cách trí tuệ nhân tạo và con người có thể thay đổi thế giới nhờ sức cộng hưởng. Trong cuốn sách mà Abraham giới thiệu trong cuộc phỏng vấn, tác giả viết rằng quân đội phải dựa vào trí tuệ nhân tạo để 'giải quyết nút thắt mang tên con người trong việc tạo mục tiêu'. Nút thắt. Tức là con người là nút thắt. Ông nói rằng dù đưa vào bao nhiêu sĩ quan tình báo đi nữa, trong thời chiến họ cũng không thể tạo đủ mục tiêu trong một ngày. Vì vậy, theo lập luận đó, cần có máy móc.

Chỉ một từ này đã chứa lối nghĩ của AI quân sự. Nút thắt. Đó là ngôn ngữ của nhà máy. Trên dây chuyền sản xuất, công đoạn chậm nhất quyết định tốc độ của toàn bộ dây chuyền. Công đoạn chậm ấy được gọi là nút thắt, và cải tiến quy trình nghĩa là loại bỏ nút thắt. Tác giả cuốn sách xem con người là phần chậm nhất trong quy trình tạo mục tiêu. Con người chậm vì nghi ngờ. Chậm vì kiểm tra. Chậm vì ngủ, vì nghỉ, vì do dự. Trong chiến tranh, sự chậm ấy bị nhìn như một khiếm khuyết. Vì vậy, ý tưởng là đặt máy móc vào vị trí chậm ấy. Nhưng điều nằm bên trong sự chậm đó chính là phán đoán. Nghi ngờ, kiểm tra, do dự không phải là tên gọi khác của sự kém hiệu quả, mà là tên gọi khác của trách nhiệm. Khi nút thắt bị loại bỏ, trách nhiệm cũng biến mất theo.

Điều đáng chú ý là trong chính cuốn sách ấy, tác giả đã vạch ra một ranh giới rõ ràng.

Abraham đã chỉ đúng đoạn này. Tác giả viết rất rõ rằng hệ thống này không được thay thế phán đoán của con người. Ông gọi đó là 'học hỏi lẫn nhau' giữa con người và trí tuệ nhân tạo. Lập trường của cuốn sách là sĩ quan tình báo mới là người nhìn kết quả rồi đưa ra quyết định, và lập trường chính thức của quân đội Israel cũng như vậy.

Một khoảng cách đã mở ra giữa lời hứa trên bản thiết kế và thực tế ngoài hiện trường.

Theo những gì Abraham nghe được từ nhiều nguồn tin, sau ngày 7 tháng 10, lời hứa ấy đã không còn được giữ, ít nhất trong một bộ phận quân đội Israel. Các sĩ quan tình báo được nói rằng, chỉ cần xác nhận mục tiêu là nam giới thì có thể chấp nhận khuyến nghị của Lavender. Không cần xem kỹ khuyến nghị đó. Không cần kiểm tra vì sao cỗ máy lại đưa ra quyết định như vậy. Con người vẫn ở trong vòng lặp. Nhưng thứ người đó thêm vào không phải là phán đoán, mà là chữ ký.

Một nguồn tin đã cô đọng tình trạng này bằng đúng một hình ảnh. Con dấu cao su. Người ấy nói mình cảm thấy như một người đóng dấu cao su lên quyết định của cỗ máy. Bề ngoài rằng con người ra quyết định vẫn còn nguyên. Có người ngồi trước màn hình, và người bấm nút phê duyệt cũng là con người. Nhưng việc người đó làm là chuẩn y một quyết định mà cỗ máy đã đưa ra rồi. Trong vòng 20 giây.

Lập trường chính thức của quân đội Israel thì khác. Trong cuộc phỏng vấn với Democracy Now, câu trả lời của quân đội Israel mà Abraham đọc lại nói rằng nhiều công cụ và phương pháp được dùng trong quá trình nhận diện mục tiêu, còn công cụ quản lý tình báo chỉ được dùng để hỗ trợ nhà phân tích. Họ nói không dùng hệ thống trí tuệ nhân tạo để nhận diện hoặc dự đoán ai là khủng bố. Hệ thống tình báo là công cụ dành cho nhà phân tích. Khoảng cách giữa câu nói ấy và lời chứng về con dấu cao su chính là sức căng của chương này. Nếu là công cụ, con người dùng công cụ. Nếu là con dấu cao su, con người bị công cụ kéo đi. Cùng một màn hình, cùng một nút bấm, cùng 20 giây ấy, một bên gọi là công cụ, bên kia gọi là người giơ tay theo lệnh. Cuốn sách này cho rằng lời chứng có trọng lượng hơn. Người dùng công cụ phải có khả năng nghi ngờ công cụ, nhưng với 20 giây và một bước xác nhận giới tính, không còn chỗ cho sự nghi ngờ chen vào.

Ở đây còn có một cỗ máy khác đẩy tốc độ lên cao hơn nữa.

Tên của nó là Fire Factory. Nhà máy lửa. Hệ thống này tự động phân bổ vũ khí. Nó tính toán mục tiêu nào dùng quả bom nào, thả bằng máy bay nào, vào lúc nào. Theo các bản tin, Fire Factory đã rút thời gian chuẩn bị tấn công từ đơn vị giờ xuống đơn vị phút. Toàn bộ quá trình tìm mục tiêu, ghép vũ khí tương ứng và chuẩn bị xuất kích đã bị nén lại.

Trong thuật ngữ quân sự, quá trình này được gọi là kill chain. Chuỗi giết chóc. Nó chỉ nhiều bước nối với nhau như một dây xích: tìm mục tiêu, cố định mục tiêu, theo dõi, ngắm bắn, tấn công và đánh giá. Khi máy móc bước vào dây xích vốn từng vận hành theo tốc độ của con người, thời gian mắc ở từng mắt xích sụp xuống. Gospel đánh dấu tòa nhà, Lavender đánh dấu con người, Where's Daddy? bắt lấy thời điểm, Fire Factory ghép vũ khí vào. Con người dùng 20 giây ở đâu đó giữa các bước ấy.

Cần nói rõ rằng vị trí của con người trong dây xích này ngày càng thu hẹp. Trong kill chain trước đây, ở mỗi bước đều có con người. Người tìm mục tiêu, người xét xem mục tiêu đó có đúng không, người chọn bom, người định thời điểm. Giữa bước này và bước kia có thời gian, và trong thời gian ấy con người có thể dừng lại, hỏi lại hoặc lọc bỏ. Khi máy móc nối các bước lại với nhau, thời gian ở giữa biến mất. Chỗ để dừng cũng biến mất theo. Dây xích chảy một mạch từ đầu đến cuối, và con người chỉ có thể đóng dấu tại một điểm nào đó trong dòng chảy. Cụm từ human in the loop là lời hứa rằng con người ở trong vòng lặp. Nhưng nếu vòng lặp quay nhanh hơn con người, thì dù ở trong vòng lặp, con người cũng không kiểm soát được vòng lặp. Ở bên trong và kiểm soát là hai việc khác nhau.

Có thể có người hỏi: nhanh thì có vấn đề gì?

Bản thân tốc độ nhanh không phải là vấn đề. Vấn đề nằm ở chỗ tốc độ ấy đã xóa mất chỗ dành cho việc xem xét. Khi con người không còn thời gian nhìn vào phán đoán của máy, họ không còn ngồi ở vị trí của người phán đoán nữa, mà ngồi ở vị trí của người cho thông qua. Có một khái niệm gọi là thiên kiến tự động hóa (automation bias). Đó là khuynh hướng con người tin một cách thiếu phê phán vào câu trả lời do máy đưa ra. Khi trên màn hình đã hiện điểm số, người bên cạnh cũng đều đang phê chuẩn, và số mục tiêu phải xử lý trong một ngày chất cao như núi, con người sẽ dễ làm theo máy hơn là nghi ngờ nó. 20 giây là đủ để thiên kiến ấy bám rễ, nhưng lại quá ngắn để sự nghi ngờ kịp chen vào.

Có thể hình dung thiên kiến tự động hóa đông cứng lại như thế nào qua một cảnh cụ thể.

Hãy thử tưởng tượng. Một sĩ quan tình báo đang ngồi trước bàn làm việc. Trên màn hình là một danh sách dài các mục tiêu cần xử lý. Hôm qua cũng vậy, hôm kia cũng vậy. Có thể ở vài trường hợp đầu tiên, ông ta đã từng nhìn vào căn cứ của điểm số. Nhưng câu trả lời do máy đưa ra nhìn chung không lệch khỏi phán đoán của những người bên cạnh. Cấp trên cũng thúc phải xử lý nhanh. Cứ như vậy, một giờ, một ngày, rồi một tuần trôi qua, bản thân việc nghi ngờ điểm số trở thành một sự kém hiệu quả. Tin vào máy thì nhẹ hơn và nhanh hơn. Nghi ngờ nuốt mất thời gian, mà thời gian lại là thứ ông ta không được phép có. Như thế, con người trượt khỏi vị trí phán đoán sang vị trí cho thông qua. Không ai ra lệnh trực tiếp, nhưng cấu trúc đã đẩy ông ta về phía đó.

Vậy trách nhiệm đã đi đâu?

Máy không chịu trách nhiệm. Máy chỉ xuất ra điểm số. Sĩ quan tình báo nói mình chỉ là một con dấu cao su. Ông ta nói mình chỉ ký vào thứ máy đã định sẵn. Bộ chỉ huy ra lệnh nói rằng con người vẫn ở trong vòng lặp và đưa ra quyết định cuối cùng. Đơn vị tạo ra hệ thống nói đó không phải là cỗ máy thay thế phán đoán, mà chỉ là một công cụ. Ai cũng nắm một phần của quyết định. Không ai nắm toàn bộ quyết định. Bom rơi xuống một ngôi nhà, cả gia đình chết, nhưng người nói cái chết ấy là quyết định của mình lại biến mất. Đây chính là khoảng trống. Chuỗi vẫn nối liền như cũ, chỉ có mắt xích mang tên trách nhiệm là rơi ra. Có lẽ nói chính xác hơn, nó không rơi ra, mà đã bị chia nhỏ qua nhiều bàn tay đến mức không còn đọng lại trong bàn tay nào.

## 4. Sinh mạng bị thay bằng thống kê: ngưỡng cho phép từ 15 đến 100

Có một thuật ngữ gọi là thiệt hại phụ.

Trong tiếng Anh là collateral damage. Đó là thuật ngữ chỉ thiệt hại đối với dân thường phát sinh cùng lúc khi tấn công một mục tiêu quân sự. Cụm từ này biến thiệt hại thành con số. Nó biến con người thành mức được phép chấp nhận. Chính chữ “phụ” đã mang ý đó. Nhánh bên, phần thêm, sản phẩm phụ không phải thứ ban đầu nhắm tới. Khi một đứa trẻ đã chết trở thành

thiệt hại phụ, cái chết ấy bị đẩy từ trung tâm của sự việc ra rìa. Ngôn ngữ trước hết đưa con người ra ngoại vi, rồi con số đặt giới hạn cho vùng ngoại vi đó. Ở Gaza, mức được phép chấp nhận ấy đã được gắn sẵn một con số.

Theo bài viết của Abraham, bộ phận luật quốc tế của quân đội Israel đã báo trước cho các sĩ quan tình báo mức thiệt hại phụ được phép đối với từng mục tiêu. Nghĩa là khi ném bom một mục tiêu cấp thấp do Lavender đánh dấu, số dân thường có thể bị giết cùng lúc đã được ấn định từ trước.

Hãy đặt các con số như những bậc thang.

Một nguồn tin tình báo nói con số ấy lên tới 20 dân thường. Cho mỗi thành viên Hamas, bất kể cấp bậc, bất kể tầm quan trọng, bất kể tuổi tác.

Một nguồn tin tình báo khác nói giới hạn đó là 15 người.

Trong quá khứ, con số ấy là 0.

Từ 0 lên 15, từ 15 lên 20. Số dân thường gắn với cái giá sinh mạng của một thành viên cấp thấp đã tăng lên như vậy. Một nguồn tin tình báo nói cả người vị thành niên cũng bị đánh dấu. Không nhiều, nhưng vì không có giới hạn tuổi, khả năng đó đã tồn tại.

Khi lên đến cấp cao, con số đổi sang một hàng chữ số khác.

Theo các nguồn tin tình báo, đối với các chỉ huy cấp cao của Hamas, ngưỡng cho phép lần đầu tiên trong lịch sử quân đội Israel đã lên tới ba chữ số. Hơn 100 người. Abraham nêu một trường hợp cụ thể. Trong cuộc tấn công nhằm vào một người từng là chỉ huy Lữ đoàn Trung tâm của Hamas, quân đội đã phê chuẩn việc giết 300 thường dân Palestine cùng với riêng người đó, theo lời một nguồn tin tình báo tham gia chiến dịch.

Hãy đặt con người trở lại vào con số 300 ấy.

300 là một con số mà nếu đếm từng người một thì khó thấy điểm kết. Dù chỉ dành 5 giây để hình dung khuôn mặt của mỗi người, cũng mất 25 phút. Nhưng trong hồ sơ mục tiêu, 300 ấy chỉ là một con số ghi trong một ô. Ô thiệt hại phụ. Trên màn hình của sĩ quan tình báo phê chuẩn vụ ném bom, 300 ấy là một dòng có thể lướt qua chỉ bằng một lần cuộn chuột. Khi con người bị đổi thành con số, chuyện như vậy trở nên có thể. 25 phút trở thành 1 giây.

Plus972 và Local Call cũng đã nói chuyện với các cư dân Palestine chứng kiến cuộc tấn công ấy. Họ nói rằng ngày hôm đó bốn tòa nhà dân cư khá lớn đã bị ném bom. Những căn hộ đầy các gia đình bị đánh bom trọn vẹn, và người bên trong chết. Abraham khẳng định đó không phải là sai sót. Con số 300 thường dân thiệt mạng ấy là con số mà quân đội Israel đã biết trước. Các nguồn tin tình báo mô tả như vậy. Một nguồn tin nói rằng trong giai đoạn đầu của thời kỳ đó, nguyên tắc tương xứng theo luật quốc tế, nếu chuyển nguyên văn cách họ nói, đã "không tồn tại".

Không tồn tại. Câu này nặng. Một sĩ quan tình báo, khi nói về hệ thống mà chính mình từng ở bên trong, đã làm chứng rằng nguyên tắc cốt lõi của luật quốc tế không vận hành. Đây không phải lời phê phán từ bên ngoài, mà là lời của người từng ở bên trong.

Nguyên tắc tương xứng chưa từng bị bãi bỏ. Nó vẫn nằm nguyên trong luật nhân đạo quốc tế. Trên giấy, nó vẫn còn sống. Nhưng tại hiện trường, nó đã ngừng vận hành. Trong hồ sơ mục tiêu, các ô như 'chỉ huy' và 'thiệt hại phụ' vẫn được điền. Hình thức vẫn đầy đủ. Abraham cho rằng chính hình thức ấy lại nguy hiểm. Tôi xin chuyển lời ông. Khi máy tạo ra hồ sơ mục tiêu, rồi trong đó các ô như chỉ huy hay thiệt hại phụ của mục tiêu được điền vào, vẻ ngoài của luật quốc tế vẫn được giữ lại. Nhưng ông nói rằng toàn bộ ý nghĩa bên trong đã biến mất.

Câu hỏi ông đặt ra xuyên suốt cả chương này.

Tôi xin chuyển nguyên lời Abraham. "Nếu thiết kế một hệ thống đánh dấu 37.000 người, xác nhận và biết rằng 10 phần trăm trong số đó thực ra không phải là lực lượng vũ trang, nhưng vẫn phê chuẩn và dùng hệ thống ấy suốt nhiều tuần mà không có sự giám sát có ý nghĩa, chẳng phải đó là vi phạm nguyên tắc phân biệt sao. Nếu phê chuẩn việc giết tới 15, 20 thường dân vì một mục tiêu mà từ góc nhìn quân sự bị xem là không quan trọng lắm, chẳng phải đó rõ ràng là vi phạm nguyên tắc tương xứng sao."

Phân biệt. Đó là nguyên tắc phải phân biệt giữa dân thường và chiến binh. Tính tương xứng. Đó là nguyên tắc thiệt hại đối với dân thường không được quá mức so với lợi ích quân sự. Hai nguyên tắc này là trụ cột của luật nhân đạo quốc tế. Abraham cho rằng các hệ thống trí tuệ nhân tạo đang rút cạn ý nghĩa khỏi những thuật ngữ ấy. Theo cách ông nói, luật quốc tế hiện đang rơi vào khủng hoảng, và những hệ thống trí tuệ nhân tạo như vậy đang làm cuộc khủng hoảng ấy sâu hơn.

Có một bối cảnh đã khiến cảnh này trở thành có thể.

Theo Washington Post của Mỹ đưa tin vào tháng 12 năm 2024, Israel đã xây dựng cái gọi là 'nhà máy trí tuệ nhân tạo' trong hơn 10 năm, và chỉ trong vài tuần đã ném bom khoảng 12.000 mục tiêu ở Gaza. Theo Plus972 và The Guardian của Anh đưa tin năm 2025, Đơn vị 8200 đã lưu trữ hàng triệu cuộc gọi của người Palestine trên đám mây Azure của công ty Mỹ Microsoft. Khối lượng ấy lên tới khoảng 11.500 terabyte. Quy mô giám sát chính là quy mô mục tiêu. Nghe càng nhiều, lưu càng nhiều, thì càng có thể chấm điểm nhiều người hơn.

Hãy thử hình dung khối lượng 11.500 terabyte. Đây là dung lượng đủ chứa hàng triệu cuộc gọi, thậm chí còn dư. Giọng nói của một người, người đó gọi cho ai và đã nói gì, được xếp chồng lên nhau trên một máy chủ nào đó ngoài Gaza. Chủ nhân của những cuộc gọi ấy không biết giọng nói của mình được lưu ở đó. Và một phần trong số những giọng nói đã được lưu ấy trở thành điểm số của Lavender. Bởi vì người đó gọi cho ai, thay điện thoại thường xuyên đến mức nào, đều là nguyên liệu tạo điểm. Giám sát là công đoạn đứng trước việc chọn mục tiêu.

Dữ liệu chất trong đám mây chảy vào máy, máy nhả ra điểm số, điểm số thành danh sách, và danh sách rơi xuống mái nhà. Việc đám mây của một công ty Mỹ nằm ở ô đầu tiên của chuỗi này cho thấy chuỗi ấy không khép lại trong phạm vi một quốc gia. Điểm này sẽ trở lại ở các chương sau.

Abraham hoài nghi trước lời nói rằng Israel sẽ tiến hành những cuộc điều tra như vậy. Ông nhắc lại hồ sơ trong quá khứ. Tôi xin chuyển lời của ông. Trong cuộc chiến Gaza năm 2014, 512 trẻ em Palestine đã chết, và Israel nói sẽ điều tra. Hàng trăm cáo buộc tội ác chiến tranh được nêu ra. Nhưng vụ duy nhất mà quân đội Israel thật sự truy tố binh sĩ lại liên quan đến việc cướp khoảng 1.000 shekel. Tất cả các vụ còn lại đều bị khép lại. Ông cho rằng lời nói cuộc điều tra đang diễn ra không có nghĩa là điều gì đó đang thay đổi. Ông nói đó là sự chế nhạo trí tuệ của chúng ta.

Đến đoạn này, cuốn sách này không thể không nghiêng về một phía.

Ghi lại lời của cả hai bên khác với việc nói rằng cả hai bên đều đúng như nhau. Quân đội Israel nói Lavender chỉ là một công cụ. Lập trường ấy đã được ghi nguyên trong phần chính. Nhưng bài giảng công khai của giám đốc Trung tâm trí tuệ nhân tạo thuộc Đơn vị 8200, cuốn sách do ông viết, lời chứng trùng khớp của sáu nguồn tin tình báo, và sự hủy diệt cả gia đình mà thống kê của Liên Hợp Quốc cho thấy. Khi đặt sức nặng này lên phía đối diện của bàn cân, bàn cân nghiêng về một phía. Chỉ riêng chữ công cụ không giải thích được 37.000 cái tên và mức cho phép mười lăm người trên mỗi người. Cuốn sách này sẽ không che giấu độ nghiêng ấy. Đồng thời, tôi cũng sẽ không xóa câu cảnh báo được ghi trong cuốn sách của người tạo ra cỗ máy, câu nói rằng không được thay thế phán đoán của con người. Bởi vì sự thật rằng ngay cả đường ranh do chính người thiết kế vạch ra cũng đã sụp đổ ngoài hiện trường, chính mâu thuẫn ấy là trọng tâm của chương này.

Lời Abraham nói ở cuối cuộc phỏng vấn hợp với phần cuối của chương này.

Ông cho rằng cuộc chiến dựa trên trí tuệ nhân tạo này nguy hiểm đối với loài người. Xin chuyển lại lời của ông. Những hệ thống như vậy cho phép con người thoát khỏi trách nhiệm. Chúng tạo hàng loạt mục tiêu ám sát tiềm tàng, hàng nghìn người, 37.000 người, nhưng vẫn giữ được vẻ ngoài tuân thủ luật pháp quốc tế. Vì đã có một cỗ máy lập hồ sơ mục tiêu. Ông nói thêm rằng, nhìn từ kinh nghiệm của các cuộc chiến trước đây, khi chiến tranh kết thúc, những hệ thống như vậy sẽ được bán cho quân đội trên khắp thế giới.

Câu nói rằng khi chiến tranh kết thúc thì chúng sẽ được bán đi là điều đáng khắc sâu.

Gaza đã trở thành một bãi thử. Lavender, Gospel, Where's Daddy, Fire Factory lần đầu tiên vận hành ở quy mô lớn tại đó. Hệ thống đã được kiểm chứng ở bãi thử sẽ trở thành hàng hóa. Nó được bán cho các quân đội khác cùng với lời quảng cáo rằng hiệu quả đã được chứng minh. Nhưng điều được gọi là kiểm chứng ở đây rốt cuộc là kiểm chứng cái gì? Đó là việc chỉ ra mục

tiêu thật nhanh. Đó là việc rút phần xem xét của một con người xuống còn 20 giây. Đó là việc phân tán trách nhiệm qua nhiều bàn tay đến mức không còn nằm lại ở bàn tay nào. Nếu thứ vượt qua bãi thử không phải là độ chính xác, mà là khoảng trống trách nhiệm này, thì thứ được bán đi cũng chính là khoảng trống ấy. Đây là lý do cuốn sách này không dừng lại ở Gaza. Cấu trúc nhìn thấy ở Gaza không phải hoàn cảnh riêng của một cuộc xung đột, mà là khuôn mẫu sẽ lặp lại trên nhiều chiến trường sau này.

Câu cuối này nối sang chương tiếp theo.

Thứ được tạo ra ở Gaza sẽ không chỉ ở lại Gaza. Cách làm chuyển việc chọn mục tiêu từ tốc độ của con người sang tốc độ của máy móc này sẽ trôi tới những cánh đồng Ukraine, rồi tới các chiến trường xa hơn nữa. 37.000 người có tên trong danh sách, 10 phần trăm trong số đó không phải là phần tử vũ trang, mức cho phép từ 15 đến 100 người cho mỗi mục tiêu. Giữa những con số ấy, điều biến mất cuối cùng chỉ là một điều. Một người có thể nói rằng cái chết ấy là trách nhiệm của mình. Khoảng trống nằm ở đó.

# Phần 1 Tốc độ của máy móc Chương 2 Ukraina: Sự tiến hóa trong thực chiến của Project Maven

## 1. Dữ liệu tràn ngập, con mắt thiếu hụt

Màn hình không dừng lại.

Một chiếc drone bay trong một giờ thì để lại một giờ video. Những chiếc drone như thế có hàng chục chiếc, có ngày lên đến hàng trăm chiếc. Trên bàn làm việc của một nhà phân tích, những hình ảnh gom lại như vậy cứ chảy mãi không dứt. Đường đất, xe tải, đàn dê, rồi lại đường đất. Ai đó phải tìm ra trong đó bàn tay đang chôn bom tự chế, bờ vai đang chuyển vũ khí, một chiếc xe dừng lại ở lối vào. Nhưng mắt người không thể giữ mãi một điểm lâu như vậy. Quân đội Mỹ đã phải trả giá đắt mới hiểu ra điều đó.

Trong cuộc chiến chống khủng bố ở Afghanistan và Iraq, các máy bay không người lái mà quân đội Mỹ cho bay đã thu về một lượng video khổng lồ mỗi giờ. Vấn đề nằm ở bước tiếp theo. Không có đủ người để xem những video ấy. Nói chính xác hơn, có người, nhưng lượng video vượt quá sức người có thể theo kịp. Trong một cuộc trò chuyện giới thiệu cuốn sách 'Project Maven' của mình, Katrina Manson đã tóm lại điểm xuất phát ấy như sau. Nước Mỹ bị "đè nặng bởi quá nhiều video drone đến mức không thể xử lý nổi", và họ muốn dùng AI để tìm ra "những thứ mà các nhà phân tích con người bị chôn vùi trong dữ liệu nên không sao phát hiện được".

Ở đây cần gặp một người. Drew Cukor. Đại tá Thủy quân lục chiến Mỹ. Tháng 10 năm 2001, hai tháng sau 9·11, ông là sĩ quan tình báo vào Afghanistan trong nhóm Thủy quân lục chiến viễn chinh đầu tiên. Theo ghi chép của Manson, ông phụ trách một ô tình báo trong các chiến dịch ban đầu cùng đồng đội Dave Spark, người mang theo một chiếc máy tính nặng nề. Và ông đang giận dữ. Ông giận điều gì. Giận vì thông tin không đến được tiền tuyến.

Cuộc chiến mà Cukor nhìn thấy là như thế này. Thông tin có ở đâu đó. Nơi ẩn náu cũ của Taliban, lối đường từng xảy ra các vụ tấn công, ai đang chế tạo loại bom nào. Nhưng thông tin ấy lại không đến được tay người đang cầm súng. Binh sĩ chết vì thiếu thông tin. Ông muốn ghi lại bom tự chế được chôn ở đâu, nổ thường xuyên ra sao, nổ trong thời tiết nào, nhưng công cụ mà quân đội Mỹ dùng khi ấy là PowerPoint và Excel. Manson nói Cukor là người làm việc giữa khoảng trống thông tin. Ông chỉ muốn có dữ liệu.

Vì vậy, hạt giống của Maven không mọc lên từ một viễn cảnh tương lai lớn lao, mà từ cơn giận của một sĩ quan tình báo. Cơn giận vì thông tin không cứu được người. Năm 2011, Cukor dồn sức đưa hệ thống tích hợp dữ liệu của Palantir vào Thủy quân lục chiến, và ông nói với Manson rằng hệ thống ấy "cứu mạng người gần như hằng đêm". Trong mắt ông, hệ thống cũ rất thô sơ. Dữ liệu bị đứt đoạn, không có quyền truy cập, biểu mẫu mỗi nơi một kiểu.

Năm 2017, cơn giận ấy tụ lại thành một dự án.

Năm đó, Bộ Quốc phòng Hoa Kỳ lập ra Algorithmic Warfare Cross-Functional Team, một nhóm liên chức năng về chiến tranh thuật toán. Tên của nó là Project Maven. Trong một căn phòng không cửa sổ ở Lầu Năm Góc, một nhóm nhỏ tụ họp lại. Phần giải thích công khai khá hẹp. Video từ drone quá nhiều, nên họ sẽ dùng AI để tự động nhận ra các vật thể trong đó, như xe cộ và con người. Mục tiêu ban đầu là Nhà nước Hồi giáo, IS. Manson nói với khán giả rằng nhiều người từng có mặt trong căn phòng ấy đang ngồi trong buổi nói chuyện về sách của ông. Họ đã ở đó.

Nhưng sau khi phỏng vấn hơn 200 người liên quan, Manson biết được một điều: tham vọng ấy ngay từ đầu đã đi xa hơn nhiều so với lời giải thích hẹp kia. Tham vọng đó đặt trên một nỗi sợ. Nỗi sợ rằng Hoa Kỳ đang tụt lại sau Trung Quốc. Khi ấy, ngân sách quốc phòng của Hoa Kỳ vẫn lớn áp đảo, nhưng Lầu Năm Góc bị ám ảnh bởi nhận thức rằng Trung Quốc đã nghiên cứu điểm yếu của Hoa Kỳ suốt 10 năm. Thứ trưởng Bộ Quốc phòng khi đó, Bob Work, tin rằng tự chủ, tức đưa con người ra khỏi chiến trường và để máy móc quyết định, là nước đi quyết định của Hoa Kỳ. Và con đường đi tới tự chủ là AI.

Theo cách nói của Cukor, đó là 'chấm trắng'. Một chấm trắng trên bản đồ. Khi bấm vào, nó trở thành tọa độ; khi đã thành tọa độ, đó là điểm có thể đưa vũ khí tới. Trong nhiều năm, ông đã mài giũa ý tưởng này bằng các bài viết. Việc nhận ra trong video có gì mới chỉ là điểm khởi đầu. Bức tranh ông vẽ là kéo thông tin xuống tận tiền tuyến, tới tận bàn tay của người đang cầm súng.

Có một vấn đề. Đó là từ 'mục tiêu'.

Một giai thoại Manson ghi trong sách cho thấy sự căng thẳng ấy. Ở một khu vực tác chiến nọ, các nhà phân tích từng phải vật lộn với hình ảnh một người nông dân. Đúng lúc quân đội Hoa Kỳ sắp tấn công, một người nông dân bước vào cánh đồng. Nhà phân tích con người mất 40 giây để tìm ra người nông dân ấy trên màn hình. Nhưng khi đưa cùng đoạn video đó chạy lại qua AI, AI lập tức nhận ra ông ta. Vì vậy đây là một ví dụ không thể tốt hơn để bênh vực AI. Rằng AI có thể cứu mạng người. Một người nào đó tại hiện trường đã viết gửi về sở chỉ huy rằng AI có thể hữu ích trong việc chọn mục tiêu và bảo vệ sinh mạng.

Câu trả lời gửi lại là thế này. "Đừng dùng từ mục tiêu. Đừng nhắc từ đó. Quốc hội đang xem xét chuyện này. Ngay trong Lầu Năm Góc cũng có người phản đối."

Vì vậy, trong một thời gian dài, từ 'mục tiêu' bị xóa khỏi ngôn ngữ chính thức của Maven. Maven là một dự án tình báo và nghiên cứu phát triển, không phải một chiến dịch tác chiến. Nhưng ít nhất trong đầu một số người, Maven luôn là việc nối thông tin với tác chiến, nối AI vào chuỗi sát thương. Cukor cho rằng quân đội Hoa Kỳ sẽ mất 20 năm để đi qua thay đổi công nghệ và văn hóa này. Theo ông, cái thật sự khó không phải là công nghệ, mà là văn hóa.

Năm 2018, tiến trình lặng lẽ ấy có một lần rung chuyển lớn.

Các nhân viên Google phát hiện công nghệ của chính công ty mình đang được dùng trong Project Maven của Lầu Năm Góc. Theo cách Manson nói trong cuộc trò chuyện, khoảng 3.000 người, còn theo các bản tin khác là khoảng 4.000 nhân viên, đã ký tên vào thư phản đối. "Chúng tôi không muốn dính chân vào ngành kinh doanh chiến tranh." Hơn mười người rời công ty. Mối lo của các nhà hoạt động nhân quyền còn đi xa hơn một bước. Đưa AI vào ở giai đoạn này rốt cuộc sẽ mở đường cho vũ khí sát thương tự chủ. Một con đường khác với lời giải thích mà Maven đưa ra.

Google quyết định không gia hạn hợp đồng. Khi đó, việc này được nhìn như một đòn chí mạng đối với nỗ lực AI của quân đội Mỹ. Nhưng kết quả Manson lần theo lại khác. Nếu hỏi Cukor, câu trả lời sẽ là "mọi chuyện đều ổn". Ý ông là các công ty khác đã ùa vào chỗ Google rút đi. Microsoft, AWS của Amazon, và Clarifai, một startup nhỏ. Matt Zeiler, người dẫn dắt Clarifai, từng là thành viên đội vô địch cuộc thi ImageNet năm 2012. Khi ấy, cách ông kiếm tiền là thị giác máy tính cho blog đám cưới. Thuật toán nhận ra mạng che mặt của cô dâu, bộ vest của chú rể, các tầng của bánh cưới. Một số nhân viên của ông thấy khó chịu khi dùng lại thứ đó cho xe tải và xe tăng trên chiến trường, và một số người đã rời đi.

Dấu vết sâu hơn mà sự rút lui của Google để lại nằm ở chỗ khác. Từ lúc đó, lập luận rằng Mỹ cần AI để đối đầu với Trung Quốc bắt đầu vang lên lớn hơn nhiều. Trong Quốc hội, trong các cuộc tranh luận công khai, câu chuyện lớn dần theo hướng: nếu Trung Quốc có AI thì Mỹ sẽ thua trong bất cứ cuộc va chạm nào. Khoảng trống mà phản ứng đạo đức khiến Lầu Năm Góc phải dè dặt đã được lấp bằng câu chuyện về mối đe dọa mang tên Trung Quốc.

Maven sống sót như vậy. Và trong một thời gian, nó không hoạt động tốt.

Các thuật toán ban đầu rất tệ. Điều Manson nói trước khán giả là chính xác. Thuật toán nhìn nhầm cây thành người, đá thành xe. Video gồm nhiều khung hình tĩnh trong một giây; ở khung này nó bắt được vật thể, sang khung khác lại bỏ sót, nên các ô vuông trên màn hình cứ chớp tắt liên tục. Những nhà phân tích đang cố tập trung lại bị phân tâm hơn. Một nhà phát triển thuật toán từng nói với Manson rằng việc nhận ra vũ khí đeo trên vai một người đôi khi phụ thuộc vào ba pixel. Chỉ ba pixel. Chính họ cũng không đủ tự tin để giao cho thuật toán một phán đoán sinh tử dựa trên khác biệt nhỏ như vậy. Cảnh Manson kể trong một cuộc phỏng vấn làm sức nặng của ba pixel ấy rõ hơn. Một người đang vác vũ khí trên vai, hay đang ôm một giỏ rau? Với thuật toán, khác biệt ấy chỉ là vài pixel. Nhưng vài pixel đó quyết định có biến một con người thành mục tiêu hay không. Vũ khí hay giỏ đi chợ trở thành phán đoán sinh tử. Ngay cả các nhà phát triển thuật toán cũng không tin thuật toán của mình đến mức giao phán đoán ấy cho máy.

Còn có một sự thật nặng nề hơn. Các thuật toán ban đầu không phân biệt được nam giới, phụ nữ và trẻ em. Loại phụ nữ và trẻ em khỏi danh sách mục tiêu là phần cốt lõi để tuân thủ luật chiến tranh, nhưng máy không làm nổi phần cốt lõi ấy. AI từng hứa hẹn nhiều điều. Nó sẽ mở rộng quy mô chiến tranh, tăng tốc độ, tăng độ chính xác. Manson cho rằng lời hứa về tốc độ

đang thực sự được giữ trong chiến dịch Iran. Nhưng lời hứa về độ tinh vi và độ chính xác thì khác, nếu nhìn vào cách thuật toán hoạt động. Ít nhất ở giai đoạn đầu, con người nhận ra tốt hơn máy rất nhiều.

Cukor gọi AI thời đó là "một túi khoai tây chiên". Đó là phép ví von Manson đã ngẫm nghĩ rất lâu. Rốt cuộc ý của nó là thế này. Bản thân thuật toán chẳng đáng bao nhiêu; điều quan trọng là hệ thống mà thuật toán ấy sẽ nằm trong đó, và cách tác chiến mà chúng ta sẽ thay đổi.

Manson kể khá kỹ về thời kỳ mọi thứ vận hành không ổn ấy. Maven thử nghiệm thuật toán ở Somalia, Afghanistan, Iraq và Qatar. Phạm vi cứ được mở rộng. Đó là chuyện vào khoảng năm 2018 đến 2019. Ngay cả lý do Microsoft tham gia cũng xuất phát từ thất bại. Vì thiếu ảnh chụp các vật thể chiến trường, Maven hỏi liệu có thể tạo dữ liệu tổng hợp để huấn luyện thuật toán hay không. Việc đó không suôn sẻ. Không hiểu sao thuật toán nhìn thấu dữ liệu tổng hợp. Tuy vậy, Microsoft trình bày thất bại ấy khéo đến mức Maven bỏ ý định dùng dữ liệu tổng hợp và quyết định dùng thuật toán của họ. Mọi chuyện diễn ra kiểu như vậy.

Đưa thuật toán vào hiện trường tác chiến còn gian nan hơn. Câu chuyện Colin Carroll mà Manson kể là như vậy. Maven có thể đưa thuật toán đầu tiên tới Somalia vì Carroll có quan hệ cũ với chỉ huy ở đó. Không phải sức mạnh công nghệ, mà là mối dây giữa con người với nhau. Năm đầu tiên sau khi được đưa ra thực địa như thế, mọi việc không suôn sẻ. Thuật toán phát hiện quá nhiều thứ rồi nhấp nháy liên tục; trên một số màn hình, thay vì khung phát hiện, nó hiện những vệt loang che mất vật đang cầm trong tay. Các nhân viên vận hành cứ thế tắt nó đi. Nhóm Maven phải cử người tới tận khu vực tác chiến để thuyết phục họ. "Xin hãy thử dùng một lần thôi, phải nghĩ đến Trung Quốc, chúng tôi biết bây giờ nó chưa ổn, nhưng chúng tôi đang chuẩn bị." Đó là chuỗi ngày đi nài nỉ.

Hệ thống ấy được đặt tên. Maven Smart System. Manson gọi nó là "Google Earth của chiến tranh". Đó là hệ thống trải chiến trường trước mắt thành bản đồ, đánh dấu lực lượng ta và địch, và khi thị giác máy tính phát hiện điều gì đó, nó chấm một điểm lên màn hình. Có người gọi nó là "Windows của chiến tranh", người khác gọi là "AI mission control", hay "hệ điều hành của chiến tranh". Maven Smart System lớn lên khi được chuyển sang National Geospatial-Intelligence Agency, NGA. Và tại đó, một bước đột phá quyết định mở ra. XVIII Airborne Corps của Lục quân bắt đầu lặp đi lặp lại các thử nghiệm gắn thuật toán vào dữ liệu vệ tinh từ năm 2020. Thứ họ muốn không phải là video từ drone. Họ muốn máy móc chỉ ra trước mắt có gì, từ dữ liệu quang điện và hồng ngoại nhìn xuống từ vệ tinh.

Trang 7 trong sách của Manson ghi lại hệ thống ấy đã đi xa đến đâu sau 10 năm. Maven Smart System được triển khai tới mọi quân chủng của quân đội Mỹ, ở nhiều nơi trên thế giới. Nó ôm vào hơn 150 luồng dữ liệu và công việc của hơn 50 công ty. Mùa xuân năm 2025, NATO bắt đầu dùng một phiên bản; đến tháng 10 cùng năm, 10 nước thành viên NATO đã xếp hàng để dùng nó cho quân đội của mình. Một nguồn tin gọi nó là "Microsoft Windows của chiến tranh".

Số tài khoản người dùng vượt 25.000. Một chấm trắng bắt đầu từ cơn giận của một sĩ quan tình báo đã trở thành hệ điều hành của chiến tranh.

Không lâu sau, thử nghiệm ấy bước lên bàn kiểm tra thật sự. Tháng 2 năm 2022, Nga xâm lược Ukraine.

## 2. Thuật toán học tuyết () chỉ sau một đêm

Đôi mắt được huấn luyện trong sa mạc đã tới một cánh đồng tuyết.

Thuật toán của Maven học từ Trung Đông. Nó học các kiểu chuyển động diễn ra ở Iraq và Afghanistan, với cát làm nền. Nhưng Ukraine có tuyết, và trên những con đường hướng tới Kyiv, xe tăng Nga xếp hàng nối nhau. Cùng là xe tăng, nhưng nền cảnh khác đi thì máy móc sẽ lẫn lộn. Manson tóm lại rất gọn: "Nó không chạy tốt."

Manson kể rằng vào khoảnh khắc ấy, quân đội Mỹ và các công ty làm thuật toán đã làm như sau. Họ điều vệ tinh. Họ chụp ảnh mới. Họ chuyển những hình ảnh mới về các hàng xe tăng cho các công ty làm thuật toán, và các công ty ấy huấn luyện lại mô hình "chỉ sau một đêm". Đôi mắt vốn đã cứng lại cho sa mạc được dạy lại suốt đêm để phù hợp với Ukraine phủ tuyết. Ngày hôm sau, năng lực phát hiện đã tiến thêm một bước.

Cảnh này nói cùng lúc hai điều. Một mặt, đó là năng lực thích nghi đáng kinh ngạc. Nhìn lại con đường công nghệ quân sự Mỹ đã đi trong 20 năm, điều ấy càng rõ. Kết cục quen thuộc thường là thế này: hứa hẹn, không chạy được, rồi dự án bị hủy. Khá hơn một chút thì có sản phẩm ra đời, nhưng hỏng nặng, người ta nói sẽ sửa, rồi cuối cùng vẫn không sửa được. Maven thì khác. Nó đến nơi trong tình trạng hỏng, nhưng họ gom được khối lượng dữ liệu học rất lớn, gắn nhãn từng dữ liệu, huấn luyện lại thuật toán, rồi triển khai lại khá nhanh. Nhờ vậy, nó trở thành một thứ có thể dùng được.

Mặt khác, chính cảnh ấy cũng phơi ra điểm yếu của công nghệ này. Đó là điều Manson và Gary Marcus, một người hoài nghi AI thẳng thắn, đã chỉ ra. Vấn đề thật sự của các hệ thống như vậy là khả năng khái quát hóa. Máy biết những gì nó đã học. Nhưng về mặt khái niệm, nó nông. Vì thế, chỉ cần chuyển sang một tình huống hơi khác, nó đã sụp. Đó cũng là lý do xe tự lái sau 30 năm, 40 năm vẫn chưa thoát khỏi giai đoạn nguyên mẫu. Nó có thể chạy ở San Francisco, nơi dữ liệu tràn ngập, nhưng sẽ dừng lại ở nơi không có dữ liệu. Ukraine phủ tuyết, với thuật toán này, là một thành phố xa lạ.

Trong câu nói huấn luyện lại qua một đêm có một rủi ro ẩn bên trong. Có thể sửa nhanh cũng có nghĩa là có thể dao động nhanh. Mô hình trôi dạt theo thời gian. Khi đầu vào thay đổi từng chút một, đầu ra cũng trượt đi. Một chuyên gia cảnh báo Manson rằng AI tạo sinh có xu hướng đồng ý với người đặt câu hỏi. Nếu hỏi: "Tôi có nên làm hành động này không? Đây có phải ý hay không?", AI dễ phụ họa: "Có." Trong giới học thuật, hiện tượng này được gọi là nịnh bợ. Trên chiến trường, một cỗ máy biết nịnh có thể khiến tình hình sôi lên mạnh hơn.

Nhưng ở Ukraine, thị giác máy tính thật sự có ích không phải chỉ nhờ riêng nó. Trong lời giải thích của Manson, điểm mấu chốt nằm ở đây. Thị giác máy tính chỉ tốt lên khi được đối chiếu chéo với tình báo tín hiệu và các nguồn tin khác. Vệ tinh chỉ ra một vật gì đó, liên lạc bị nghe lén củng cố điều ấy, rồi một nguồn khác xác nhận thêm lần nữa. Chỉ khi các lớp thông tin chồng lên nhau, người ta mới nắm được đó là gì. Manson kể rằng trong năm đầu tiên ấy, quân đội Mỹ đã thật sự trở nên rất giỏi việc này, ít nhất là theo cách họ nhìn nhận.

Niềm tin bắt đầu lớn lên. Theo Manson, niềm tin chính là một trục khác trong việc thử nghiệm và nuôi lớn AI. Nó lớn đến mức nào. Trong một trường hợp, quân đội Ukraine chỉ dựa vào các nguồn tin của mình thì không biết đang đánh vào thứ gì. Nhưng phía Mỹ có thể nói: "Hãy tin chúng tôi và đánh đi." Bởi thứ trong mắt quân Ukraine chỉ là một hình chữ nhật, hoặc một chiếc xe tải, lại hiện ra trong mắt quân đội Mỹ như một thứ khác. Manson kể rằng đó là một bệ phóng di động.

Quân đội Mỹ không phải bên trực tiếp tham chiến ở Ukraine. Vì thế, họ cần một thứ ngôn ngữ mới. Không phải 'mục tiêu', mà là 'điểm quan tâm', tiếng Anh là points of interest. Manson nói đó là một cụm từ "gần như được phát minh cho chính việc này". Mọi thứ cho đến ngay trước quyết định khai hỏa. Mỹ không chỉ thị cho Ukraine phải đánh cái gì. Mỹ cũng không chuyển giao tin tình báo mật của mình. Họ chỉ dùng thị giác máy tính để chỉ ra vật thể nằm ở đâu, rồi chia sẻ vị trí đó với Ukraine. Việc biến nó thành mục tiêu là phần việc của Ukraine. Manson viết rằng gần như mọi nguồn tin của ông, vào một lúc nào đó, đã lỡ miệng gọi nó là "mục tiêu". Nhưng họ cẩn thận không gọi nó là 'mục tiêu hợp pháp'.

Đó là một cuộc đi dây bằng ngôn từ. Chỉ một từ cũng có thể quyết định có trở thành bên trực tiếp tham chiến hay không. Nhưng điều diễn ra trên màn hình, dù dùng từ nào, vẫn là việc một điểm biến thành tọa độ, rồi tọa độ gọi vũ khí đến. Khoảng trống nằm ngay trong cuộc đi dây ấy. Ai đã biến điểm đó thành mục tiêu. Thuật toán đặt điểm, nước Mỹ chuyển điểm, hay Ukraine bóp cò. Trách nhiệm bị tản ra giữa ba bên.

Cũng có những vấn đề không thể giải bằng công nghệ. Manson thấy lạ việc Maven không hiện lên trơn tru ở Bộ Tư lệnh châu Âu. Mạng quá phức tạp. Các gói dữ liệu phải băng qua Đại Tây Dương hai lần, có lúc bốn lần. Trên đường đi, gói dữ liệu bị mất hoặc chậm lại. Muốn vận hành hệ thống mật thì cần thiết bị mã hóa, và chính thứ đó trở thành nút nghẽn. Muốn chuyển một thiết bị mã hóa lớn hơn, phải gọi điện xin phép một nhân vật rất cao cấp trong cộng đồng tình báo. Lý do là một thiết bị đã được Cơ quan An ninh Quốc gia Mỹ chứng nhận. Mọi thứ phải khớp với nhau, nhưng chúng đã không khớp. Nỗ lực tăng năng lực xử lý của AI lập tức đánh thức tất cả các nút nghẽn khác đang nằm ẩn bên cạnh nó.

Dù vậy, kết quả rất rõ. Đến đầu năm 2024, Ukraine đã phá hủy hơn 2.600 xe tăng Nga. Xe bọc thép thì gần 5.000 chiếc. Ở một thời điểm mà Manson ghi trong sách, số điểm quan tâm mà Mỹ chuyển cho Ukraine có lúc lên tới tối đa 267 điểm mỗi ngày. Người dẫn cuộc trò chuyện,

Gregory Allen, nói đó là con số chưa từng có trong lịch sử chia sẻ tình báo.

Nhưng Manson nhìn cùng câu chuyện ấy chậm lại một nhịp. Sự hợp tác đó bắt đầu chậm đi từ năm sau. Manson xem đây không hẳn là câu chuyện về công nghệ, mà là câu chuyện về chính sách và niềm tin. Nhóm đầu tiên đã xây được với Ukraine một quan hệ đủ gần để nói rằng: "Hãy tin chúng tôi và đánh đi." Có những việc có thể không qua được thước đo chính sách, nhưng dưới sức ép của chiến tranh, họ cứ làm. Nhóm mới đến không có mức thân thiết ấy. Tranh luận nổ ra quanh việc liệu chuyển các điểm quan tâm như vậy có ổn không, hệ thống mật của Mỹ có bị phơi ra trước rủi ro không. Rồi đến một lúc, quan hệ ấy không còn liền mạch nữa. Ukraine tìm một con đường khác. Bằng drone của chính mình.

Bản thân dữ liệu cũng có lúc lung lay. Manson nêu một trường hợp dữ liệu huấn luyện bị hỏng trong giai đoạn đầu của Maven. Không phải do kẻ thù bên ngoài, mà do những người gắn nhãn dữ liệu đã kiệt sức vì công việc, họ điền đầy lời chửi rủa vào dữ liệu huấn luyện và không hợp tác. Nếu cứ tiếp tục huấn luyện bằng loại dữ liệu đó, chất lượng thuật toán hẳn sẽ bị ảnh hưởng. Đưa dữ liệu được gắn nhãn đúng vào hệ thống, rồi kiểm tra chéo với các nguồn khác. Việc trông có vẻ bình thường ấy thật ra lại là việc khó nhất.

Sự đổi ý của một người trong giai đoạn này cho thấy số phận của Maven. Đô đốc Whitworth. Ông từng là người nghi ngờ Maven gay gắt hơn ai hết. Là một nhân vật tình báo đã làm việc lâu năm trong chu trình nhắm mục tiêu, ông lo về trách nhiệm. Khi AI góp phần gây ra sai sót mục tiêu, ai sẽ đứng trước Quốc hội để bảo vệ chuyện đó. Liệu hệ thống có đang đi đường tắt, bỏ qua chu trình nhắm mục tiêu hay không. Khi ông tiếp quản NGA và phải gánh Maven, những người của Maven lo rằng ông sẽ khai tử nó. Nhưng điều khiến ông quay lại, theo chính lời ông nói với Manson, là sự thật rằng Maven tự cập nhật và phản hồi theo thực tế chiến tranh nhanh hơn bất cứ thứ gì ông từng thấy. Tính linh hoạt ấy của phần mềm đã biến một người hoài nghi thành người tin theo. Ông cố gắng mô tả rõ khi nào mô hình thành công, khi nào thất bại, rồi báo cho người dùng biết. Có người gọi đó là quá trình Maven 'trưởng thành'. Dẫu vậy, từ khoảng thời gian ấy, những lời phàn nàn rằng Maven chậm lại cũng bắt đầu đi kèm.

Máy móc học chỉ cần một đêm. Niềm tin giữa con người với nhau thì không chuyển giao nhanh như vậy.

## 3. Kill chain rút xuống còn 18 phút

Một bệ phóng cơ động biến mất chỉ 18 phút sau khi bị phát hiện.

Manson ghi lại cảnh này trong sách, và trong cuộc trò chuyện, người dẫn lại nhắc đến con số ấy. AI của Mỹ đã phát hiện một bệ phóng tên lửa cơ động, tiếng Anh gọi là transporter erector launcher, và từ lúc phát hiện đến lúc phá hủy mất 18 phút. Một bệ phóng tên lửa được chở trên xe tải, loại mục tiêu mà trong quân đội ai cũng biết là rất khó bắt. Vậy mà chỉ 18 phút.

Nên hiểu con số này như thế nào? Trước hết, cần nói rõ nguồn. Con số 18 phút, cả chi tiết về việc huấn luyện lại mô hình nhận diện tuyết chỉ trong một đêm, đều dựa trên bài viết của Katrina Manson và cuốn sách Project Maven của bà. Manson nhiều lần nói rằng phần lớn ví dụ trong sách của bà không phải là chuyện thuật toán thành công, mà là chuyện thuật toán thất bại, rồi học dần từ chính thất bại ấy. Vì vậy, 18 phút không phải là lời khoe khoang. Đó là một ghi chép. Một ghi chép cho thấy thứ gì đã trở nên nhanh hơn.

Trong quân đội có một tên gọi cho chuỗi này. Kill chain, tức chuỗi sát thương. Nói đầy đủ hơn, đó là các bước phát hiện, cố định mục tiêu, theo dõi, nhắm bắn, giao chiến và đánh giá. Theo chữ cái đầu trong tiếng Anh, nó còn được gọi là F2T2EA. Find, fix, track, target, engage, assess: tìm, ghim, bám theo, nhắm, đánh, rồi xem xét. Khi con người vận hành chuỗi này bằng tay, thời gian được tính bằng giờ. Khi đưa AI vào, thời gian ấy rút xuống còn tính bằng phút. 18 phút là một lát cắt của sự nén thời gian đó.

Quy mô của sự nén ấy hiện rõ hơn trong một con số khác. Manson ghi lại ở trang 7 của cuốn sách một con số bà nghe từ một quan chức NGA, rồi đọc nguyên văn con số đó trong một cuộc trò chuyện.

Nhờ thị giác máy tính, số mục tiêu mà Mỹ có thể tấn công trong một ngày đã tăng từ dưới 100 lên 1.000.

Khi mô hình ngôn ngữ lớn, tức LLM, được tích hợp vào nền tảng Maven, con số ấy tăng gấp năm lần, lên 5.000 mục tiêu mỗi ngày.

Từ dưới 100, lên 1.000, rồi lên 5.000. Nếu nhìn từng nấc một, ta thấy chuyện gì đã xảy ra. Việc tạo ra mục tiêu đã chuyển từ tốc độ của con người sang tốc độ của máy.

Đó là điều các nhà chiến lược quân sự gọi là nén kill chain. Cùng một chuỗi ấy, trong cùng một khoảng thời gian, được quay nhiều lần hơn. Khi máy nén công việc của một phân tích viên, người ta bắt đầu nói rằng một đơn vị với vài chục người có thể làm việc mà trước đây cần đến hàng nghìn phân tích viên. Một viện nghiên cứu quân sự của Mỹ, trong đánh giá về tự chủ và AI, đã nêu tốc độ, độ chính xác, phối hợp và tầm với, duy trì và sát thương, cùng sức bền là những năng lực cốt lõi mà AI hứa hẹn đem lại. Riêng về tốc độ, báo cáo viết rất rõ. Tốc độ của chiến tranh hiện đại đã vượt quá tốc độ con người có thể định hướng, hiểu tình hình và hành động. Khi có thể thua ở tốc độ của máy, chi phí xâm lược tăng lên, và vì thế năng lực răn đe mạnh hơn. Nhưng chính báo cáo ấy cũng tự kéo tay mình lại. Tốc độ của máy dễ dẫn tới nhận nhầm và phán đoán sai; nếu vòng OODA quay nhanh đến mức con người không thể can thiệp, ranh giới giữa tự chủ có giám sát và tự chủ hoàn toàn sẽ mờ đi, thậm chí mất ý nghĩa. Trong một cuộc war game, một hệ thống phòng không đặt ở chế độ tự chủ không chỉ đánh chặn tên lửa, mà còn phóng đòn phản kích về phía Triều Tiên, khiến khủng hoảng leo thang. Một câu trong báo cáo ấy xuyên suốt chương này. Tốc độ của máy sẽ mãi lệch khỏi sự kiểm soát có ý nghĩa của con người.

Nhưng LLM đã làm chính xác việc gì? Câu trả lời của Manson bất ngờ là khá giản dị. LLM không làm thị giác máy tính. Việc chọn ra đâu là mục tiêu không thuộc phần việc của LLM. Việc LLM làm là tóm tắt báo cáo tình báo và đẩy nhanh thủ tục cần cho phê duyệt mục tiêu. Nó không thay con người phê duyệt. Nó không đóng dấu thay con người. Nó chỉ xử lý thay những việc hành chính đưa hồ sơ qua lại trước khi được phê duyệt, những việc lặt vặt để hồ sơ đi qua khâu rà soát pháp lý. Manson gọi đó là phần 'hành chính' của chu kỳ nhắm mục tiêu. Không phải phần nhận diện.

Ở đây, lời chứng của một người trở nên quan trọng. Một chuẩn úy mà Manson gặp ban đầu là người hoài nghi. Ông không muốn thêm một công cụ nhắm mục tiêu nào nữa, nên đã chửi mắng và đuổi nhân viên Palantir đến trình diễn hệ thống ra ngoài. Sau cuộc rút quân khỏi Afghanistan, ông đổi ý. Điều ông nhìn thấy không phải bản thân AI, mà là sức mạnh của việc hợp nhất dữ liệu. Sức mạnh gom mọi thứ về một nơi. Ông xin lỗi nhân viên Palantir và bắt đầu dùng hệ thống. Ông nói mình mất ba ngày để dùng thành thạo. Và thao tác mà ông trực tiếp trình diễn Maven Smart System cho Manson, rồi mô tả lại, là một cảnh sẽ còn đọng lại mãi trong chương này.

Chấp nhận, chấp nhận, chấp nhận.

Khi AI hiện kết quả phát hiện, ông bấm 'có, có, có'. Hệ thống vốn được thiết kế để vận hành như vậy. Chính viên chuẩn úy ấy cũng biết rủi ro. Rủi ro về thiên kiến tự động hóa. Ở tốc độ cao, người ta không thể nhìn thấy chuyện gì đang diễn ra bên dưới nắp máy. Mỗi lần đều phải cân nhắc nên soi kỹ một đề xuất đến mức nào, kiểm tra chéo ra sao, có thể tin hệ thống đến đâu. Nhưng khi đã cuốn theo nhịp 'có, có, có', sự cân nhắc ấy mờ đi.

Manson nắm bắt sự thu hẹp đó bằng con số. Trong cách Quân đoàn Dù 18 dùng hệ thống, có sáu vị trí mà con người can dự vào quyết định trong chu trình nhắm mục tiêu. Nhờ AI, con người được rút khỏi bốn trong sáu vị trí ấy. Chỉ còn lại hai. Họ nói quyết định vẫn thuộc về con người. Nhưng việc kiểm tra dữ liệu, những thủ tục kiểu đó, đã dần dần chuyển sang tự động. Theo một cách nói khác trong quân đội, con người ở 'trên vòng lặp'. Nghĩa là họ đứng nhìn như người giám sát để có thể dừng lại nếu có gì đó đi sai. Điều này khác với việc mỗi lần đều bấm 'thực thi'.

Từ sáu xuống hai. Đọc sự thu hẹp này theo hướng nào chính là câu hỏi của cả cuốn sách. Nếu đọc theo hiệu suất, đó là tiến bộ; nếu đọc theo trách nhiệm, đó là khoảng trống.

Gary Marcus giải thích điểm này bằng tâm lý học nhận thức. Ngay cả khi đặt con người trong vòng lặp, nếu ép người đó quá mạnh, trên thực tế họ chẳng khác gì đang ở ngoài vòng lặp. Con người không thể duy trì sự cảnh giác trong thời gian dài. Nghiên cứu gốc dẫn đến phát hiện này được thực hiện với lính vận hành radar. Nếu buộc một người đánh giá 80 mục tiêu trong một giờ, người đó không thể làm tốt việc ấy. Vì vậy họ bắt đầu tin AI. Marcus gọi đó là 'hiệu ứng trông có vẻ ổn với tôi'. Khi một đoạn văn đúng ngữ pháp và trôi chảy, người ta sẽ nói 'trông có

vẻ ổn', rồi không nhận ra lỗi nằm bên trong.

Khoảng trống ở đây lại mở rộng thêm một lần nữa. Người ta nói con người vẫn ở trong vòng lặp. Nhưng nếu người đó thông qua hàng chục trường hợp mỗi giờ bằng 'có, có, có', thì ông ta đang ở trong vòng lặp hay ngoài vòng lặp? Về hình thức, ông ta ở trong. Về thực chất, ông ta ở ngoài. Trách nhiệm thoát ra qua khoảng cách giữa hình thức và thực chất ấy.

Có một sự thật khác mà Manson bám đến cùng. AI có thể chọn mục tiêu, nhưng cũng có thể trở thành chiếc lá vả che đậy. Một chỉ huy có thể định trước rằng 'tôi sẽ chấp nhận mức thiệt hại phụ này', rồi giao việc chọn mục tiêu cho AI. Khi đó, vấn đề thật sự không phải AI chính xác đến đâu. Vấn đề thật sự là quyết định ban đầu ấy: sẽ cho phép bao nhiêu thiệt hại đối với thường dân. Manson nói luôn có hai câu hỏi được đặt ra cùng lúc. Một là độ chính xác của hệ thống, chi phí của sai sót. Liệu có đang bỏ sót một ngôi trường nào không. Câu kia là liệu công nghệ này có đang được dùng để đẩy trách nhiệm sang máy móc hay không.

Nguy cơ của chiếc 'lá vả' ấy không phải là chuyện trừu tượng. Manson đưa ra một trường hợp cho thấy AI chen vào việc tính toán thiệt hại phụ như thế nào. Đêm đó năm 2024, khi quân đội Mỹ trả đũa cái chết của ba binh sĩ bằng cách tấn công 85 mục tiêu ở Iraq và Syria, AI của Maven đã được dùng để thu hẹp danh sách mục tiêu ấy. Nhưng Airwars, một tổ chức phi chính phủ theo dõi thiệt hại dân sự, đưa tin rằng trong số những người chết đêm đó có một dân thường. Họ gọi đây là 'thương vong dân sự đầu tiên có AI hỗ trợ'. Chính Manson không tự mình đưa ra tuyên bố ấy trong sách. Ông chỉ đặt câu hỏi. Khi đưa AI vào hệ thống, thiệt hại phụ sẽ thay đổi ra sao.

Một cái bóng tương tự cũng phủ lên một sự kiện sớm hơn. Năm 2019, trong chiến dịch mà Tổng thống Trump công bố việc tiêu diệt Baghdadi, Manson viết trong sách rằng AI đang theo dõi một chiếc xe van màu trắng và quan sát nhiều xe khác. Chiếc xe van trắng ấy về sau gây tranh cãi. Đài phát thanh công cộng NPR của Mỹ đưa tin rằng trong xe có dân thường, một người bị thương nặng và một người thiệt mạng. Điều đó không có nghĩa AI trực tiếp chịu trách nhiệm cho cuộc tấn công ấy. Chỉ là AI có thể đã báo cho người vận hành biết sự hiện diện của chiếc xe, và người vận hành đã đánh giá nó là mối đe dọa. Cuối cùng, không có kết luận công khai nào về việc trên xe có dân thường hay không.

Lịch sử phủ bóng lên nguy cơ thứ hai này. Trong sách, Manson xem xét vụ Mỹ đánh nhầm đại sứ quán Trung Quốc tại Belgrade năm 1999. Danh sách lỗi được công bố về sau khá đáng chú ý. Đại sứ quán nằm gần mục tiêu thật; trên một số bản đồ nó bị ghi sai, nhưng trên các bản đồ khác lại được ghi đúng. Một việc rất thực tế, gọi là cơ sở dữ liệu. Nền tảng ấy đã không được giữ ở trạng thái mới nhất. Ngay cả khi có người cảm thấy có gì đó sai, người ấy cũng không thể gọi điện để truyền đạt nghi ngờ đó. Không tìm được đúng người.

Câu hỏi Manson đặt ra lạnh lùng. Nếu đan các hệ thống này thật chặt, có thể lấp được khoảng cách an toàn ấy. Nhưng trong thời đại tác chiến cực nhanh, chính thời gian dành cho

liên lạc và xác nhận lại trở thành yếu tố quyết định. Có một câu nói cũ: "Đưa rác vào thì rác đi ra". Dữ liệu cũ. Câu hỏi là liệu hệ thống có đủ tự biết mình để yêu cầu dữ liệu mới hay không. Chuỗi hành động đã rút xuống còn 18 phút cũng đặt thêm một câu hỏi: trong 18 phút ấy, còn lại bao nhiêu thời gian để nhận ra lỗi.

Và Manson làm rõ một điều. Ở Ukraine, và trong trường hợp mà Bộ Tư lệnh Trung tâm Mỹ lần đầu công khai xác nhận vào tháng 2 năm 2024, con người vẫn ở trong vòng kiểm soát. Giám đốc công nghệ của Bộ Tư lệnh Trung tâm nói với Manson rằng khi quân đội Mỹ trả đũa cái chết của ba binh sĩ bằng cách tấn công 85 mục tiêu ở Iraq và Syria, AI của Maven đã được dùng để thu hẹp xuống 85 mục tiêu ấy. Khi được hỏi liệu con người có cùng tham gia hay không, Manson trả lời: "Tất nhiên, tôi muốn nói rõ điều đó". Vũ khí sát thương hoàn toàn tự chủ mà các nhà hoạt động lo sợ vẫn chưa tồn tại ở quy mô họ lo ngại.

Tuy vậy, Manson không xem nhẹ chữ "chưa" ấy. Ông khẳng định rằng điều đó không có nghĩa Mỹ không tìm cách phát triển nó. Ông nói Lầu Năm Góc đã có chính sách hậu thuẫn vũ khí hoàn toàn tự chủ, và trong sách ông đã tìm hiểu cách Lầu Năm Góc đi theo con đường đó. Ở cuối con đường ấy, ông cũng ghi rằng nỗ lực đưa AI lên drone để tự động nhận diện mục tiêu đang vấp váp.

Tại đây xuất hiện lời của một người mà Manson để lại trong đoạn cuối cuốn sách. Ông hỏi Cukor về lập luận quen thuộc rằng công nghệ và chính sách là hai thứ hoàn toàn tách biệt. Chẳng phải rốt cuộc chính công nghệ ấy mới cho phép mức độ sử dụng cao hay sao. Cukor thừa nhận: "Đúng". Và cuối năm ngoái, ông đã nói với Manson như sau.

"Hãy nhìn thẳng vào chính mình trong gương và kiểm tra xem chúng ta có thận trọng hay không. Chúng ta có tất cả những công nghệ này. Chúng ta có phải là người quản lý tốt nhất của chúng không?"

Đó là một câu kết thúc bằng câu hỏi. Nó không đưa ra câu trả lời.

Jack Shanahan, cấp trên cũ của Cukor và là người phụ trách đầu tiên của Maven, đã trả lời cùng câu hỏi ấy từ một góc khác. Lời ông được Manson trích dẫn như sau: "Không một LLM nào ở hình thức hiện nay nên được xem là ứng viên để dùng trong một hệ thống vũ khí tự động hoàn toàn có khả năng gây chết người. Bản thân việc đưa ra đề xuất như vậy đã là phi lý." Ngay cả những người đã nuôi lớn công nghệ ấy ở khoảng cách gần nhất cũng vạch một ranh giới trước việc để cỗ máy đứng một mình ở cuối chuỗi đó. Có lẽ không cố hàn gắn mâu thuẫn ấy là tất cả những gì chương này có thể làm.

Manson nói rằng khi viết cuốn sách, ông thường nghĩ đến một việc. Các phóng viên thường đi cùng các đơn vị quân đội Mỹ để tác nghiệp. Họ chứng kiến bên cạnh mình cảnh binh sĩ đặt mạng sống vào cuộc. Ông nói mình muốn viết một cuốn sách trong đó độc giả không đi cùng con người, mà đi cùng thuật toán. Một cuốn sách bước vào chiến tranh cùng thuật toán, nhìn

vào việc đưa AI ra tiền tuyến thật sự có nghĩa là gì. Ông nói mình chưa đi đến tận cùng việc đó. Nhưng nỗ lực còn dang dở ấy cũng chính là con đường mà chương này đã lần theo.

Màn hình vẫn không dừng lại. Chỉ là giờ đây, trên màn hình ấy, thời gian để một điểm biến thành tọa độ, rồi tọa độ gọi vũ khí tới, là 18 phút. Chỗ để con người chen vào quyết định đã giảm từ sáu xuống còn hai. Và trong 18 phút ấy, ở hai chỗ còn lại ấy, vị trí của câu hỏi ai đã bóp cò đang dần trống đi. Ai sẽ lấp vào chỗ trống đó, đến nay vẫn chưa ai trả lời được.

# Phần 2 Vượt qua ranh giới kiểm soát Chương 3 Iran: Epic Fury và giết chóc theo tốc độ của máy móc

## 1. 60 giây: Khamenei và 1.000 mục tiêu trong 24 giờ

Sáng ngày 28 tháng 2 năm 2026. Trong một ngôi nhà an toàn ở phía bắc Tehran, một ông lão 86 tuổi đang uống trà. Ayatollah Ali Khamenei, Lãnh tụ Tối cao Iran. Ông là người đã nắm trong tay tôn giáo, quân đội và ngoại giao của một quốc gia trong gần 37 năm. Hôm ấy, không chỉ các vệ sĩ của ông biết vị trí của ông.

Trên một màn hình ở Washington, ngôi nhà an toàn của ông hiện lên như một tọa độ.

Việc mà Hoa Kỳ đặt tên là 'Operation Epic Fury' đã kết thúc chỉ trong 60 giây. Theo tin của AP và Military Times, ba mục tiêu bị tấn công cùng một thời điểm. Gần như đồng thời. Không phải một nơi trúng đòn rồi đến nơi tiếp theo, mà ba nơi sụp xuống trong cùng một nhịp thở. Các bản tin ấy cho biết máy bay ném bom tàng hình B-2 đã thả bom nặng 2.000 pound. Khamenei và khoảng 40 quan chức cấp cao thiệt mạng trong một phút đó.

Trong một chiến dịch quân sự, 1 phút thường là khoảng thời gian quá ngắn để bắt đầu bất cứ việc gì. Một chiếc máy bay rời khỏi đường băng còn mất lâu hơn thế. Vậy mà phần lớn ban lãnh đạo của một quốc gia đã biến mất trong đúng 1 phút ấy.

Hãy nghĩ lại câu nói ba nơi sụp xuống cùng lúc. Nếu đánh ba nơi theo thứ tự, ngay khi đòn tấn công đầu tiên xảy ra, người ở hai nơi còn lại sẽ có thời gian chạy thoát. Tiếng nổ vang lên, còi báo động rú lên, vệ sĩ đưa lãnh đạo rời khỏi chỗ đó. Vì vậy, chiến dịch chặt đầu luôn thất bại nếu đòn đầu tiên trượt mục tiêu. Bởi mục tiêu sẽ tản ra. Tấn công đồng thời có nghĩa là không cho họ khoảng trống để tản ra. Muốn ba nơi cùng sụp xuống trong cùng 1 phút, vị trí của cả ba nơi phải được xác định ở cùng một thời điểm, và ba đòn đánh phải sẵn sàng ở cùng một thời điểm. Gần như không thể dùng tay người để khớp thông tin của ba nơi cùng lúc. Chính tính đồng thời này là nơi máy móc bước vào.

Khi lần theo lý do khiến chiến dịch này có thể diễn ra, ta sẽ chạm tới một câu hỏi. Con người có thể ra quyết định nhanh đến vậy không.

Câu trả lời là: nếu chỉ có con người thì không.

Tổng thống Hoa Kỳ phê chuẩn chiến dịch vào ngày 27 tháng 2. Chiến dịch bắt đầu vào buổi sáng ngày 28 hôm sau, theo giờ chuẩn Iran. Nhiều cơ quan truyền thông cho biết Cơ quan Tình báo Trung ương Hoa Kỳ (CIA) đã theo dõi lịch trình di chuyển của giới lãnh đạo Iran trong nhiều tháng. Nhiều tháng theo dõi và 60 giây tấn công. Khoảng cách giữa hai việc ấy tạo nên sức căng của cả chương này. Thông tin được tích lũy chậm rãi, còn quyết định rơi xuống trong khoảnh khắc. Ở giữa hai đầu ấy, khoảng trống để con người dừng lại và cân nhắc ngày càng hẹp. Israel

cùng thực hiện chiến dịch này dưới tên gọi 'Roaring Lion'. Một bên là Epic Fury, bên kia là Roaring Lion. Hai cái tên của cùng một đêm.

Nhưng giữa theo dõi và tấn công luôn có đoạn khó nhất. Đó là biến thông tin thu thập được thành mục tiêu. Đánh vào đâu. Đánh ai. Tọa độ này lúc này có đúng không. Theo cách truyền thống, con người làm việc đó. Các nhà phân tích xem ảnh vệ tinh, nghe thông tin liên lạc, đặt nhiều nguồn tin chồng lên nhau, rồi mất nhiều ngày để cân nhắc xem đây có đúng là mục tiêu hay không.

Lần này, máy móc đã tham gia hỗ trợ việc đó.

Theo các bản tin của Bloomberg và Democracy Now, hệ thống do Palantir vận hành đã được dùng trong việc chọn mục tiêu cho chiến dịch này. Đó là hệ thống phân tích mục tiêu mà quân đội Hoa Kỳ gọi là 'Project Maven' hoặc 'Maven Smart System'. Bloomberg cho biết bên trong hệ thống ấy có tích hợp Claude, trí tuệ nhân tạo của Anthropic. Nói cách khác, một chatbot do công ty Hoa Kỳ tạo ra, được đặt trên một nền tảng phân tích tình báo cũng do công ty Hoa Kỳ tạo ra, đã chạm tới việc nhắm vào giới lãnh đạo của một quốc gia.

Nhìn vào các con số sẽ thấy quy mô của sự thay đổi này.

Theo Bloomberg, trong 24 giờ đầu của chiến dịch, quân đội Hoa Kỳ đã xử lý 1.000 mục tiêu.

Sau 3 tuần, con số ấy tăng lên khoảng từ 5.500 đến 6.000 mục tiêu.

Đến ngày 6 tháng 4, con số này đạt khoảng 13.000 mục tiêu.

Một nghìn mục tiêu mỗi ngày. Đó là tốc độ mà bàn tay con người gần như không thể chạm tới. Bloomberg cho biết nếu đây là Chiến tranh Iraq năm 2003, để phân tích số mục tiêu cùng quy mô, cần khoảng 2.000 chuyên viên phân tích bám vào công việc đó. Lần này, khoảng 20 người đã làm việc ấy. Nói rằng công việc của 2.000 người được 20 người làm không có nghĩa là con người đã thông minh hơn gấp trăm lần. Nó có nghĩa là máy móc đã chen vào giữa con người và mục tiêu, rồi nén thời gian ở khoảng giữa ấy lại.

Craig Jones, học giả luật của Đại học Newcastle, đã tóm tắt thay đổi này trong một câu trên Democracy Now. Trí tuệ nhân tạo đã "nén hàng chục nghìn giờ lao động của con người xuống còn vài giây, vài phút". Những phán đoán từng mất vài ngày nay kết thúc trong vài phút; những kiểm chứng từng mất vài giờ nay trôi qua trong vài giây.

Muốn thấy sự nén này có nghĩa gì, chỉ cần nhìn vào bên trong công việc chọn mục tiêu. Quân đội có một quy trình để xác định mục tiêu. Trong tiếng Anh gọi là 'kill chain', có thể hiểu là chuỗi chết chóc. Đó là chuỗi bước gồm tìm xem sẽ đánh vào đâu, cố định vị trí của nó, xác nhận xem có đúng không, rồi cuối cùng tấn công. Tìm, cố định, kết thúc. Theo cách truyền thống, ở từng mắt xích của chuỗi này đều có con người ngồi đó. Người xem ảnh vệ tinh, người nghe tín hiệu liên lạc, sĩ quan pháp lý, chỉ huy. Mỗi người dừng lại một lần ở mắt xích của mình và hỏi: "Có

đúng không?"

Khi máy móc bước vào, những khoảng dừng ấy biến mất. Nói chính xác hơn, chúng ngắn lại. Vì máy xử lý cùng lúc nhiều mắt xích trong chuỗi, chỗ để con người dừng lại ít đi. Việc tìm kiếm, cố định và xác nhận gần như diễn ra đồng thời trên cùng một màn hình. Vì vậy 1.000 mục tiêu có thể chảy qua trong vòng 24 giờ. Không phải chuỗi tự nó nhanh hơn; nó nhanh hơn đúng bằng phần con người bị rút ra khỏi chuỗi.

Ở đây còn có thêm một yếu tố tâm lý chen vào. Đó là thiên kiến tự động hóa. Khi máy đưa ra đáp án, con người có xu hướng chấp nhận đáp án ấy hơn là nghi ngờ nó. Giống như chúng ta thường không tính lại con số mà máy tính bỏ túi đưa ra. Khi tọa độ hiện lên trên màn hình mục tiêu và có dòng chữ "đây là mục tiêu", người chỉ có 1 phút sẽ dễ ký xác nhận hơn là lật ngược kết luận ấy. Niềm tin vào máy làm mềm đi phán đoán cuối cùng của con người. Thiên kiến này quan trọng đến mức sẽ được bàn riêng ở phần sau của cuốn sách. Lúc này chỉ cần nhớ một điều. Việc con người có mặt trong chuỗi và việc con người thật sự đang phán đoán không phải là cùng một chuyện.

Đến đây cần dừng lại một chút. Tốc độ nhanh hơn cũng có thể là điều tốt. Thông tin nhanh có thể kiềm chế chiến tranh. Nếu biết trước đối phương đang làm gì, ta cũng có thể tránh va chạm. Dario Amodei, tổng giám đốc công ty tạo ra Claude, cũng nói với Bloomberg theo hướng ấy. Năng lực thông tin vượt trội có thể ngăn xung đột. Ông cũng nói rằng ông ủng hộ việc Hoa Kỳ trở thành một chủ thể mạnh hơn. Ông tự gọi mình là người yêu nước.

Nhưng tốc độ có một gương mặt khác.

Phán đoán càng nhanh, thời gian để nhìn kỹ vào phán đoán ấy càng ít. Nếu xử lý 1.000 mục tiêu trong vòng 24 giờ, thì về mặt số học, con người chỉ có hơn 1 phút cho mỗi mục tiêu. Trong 1 phút ấy, con người có thể xem xét được gì. Tọa độ có đúng không, tòa nhà đó hiện vẫn là cơ sở quân sự không, bên trong có ai, liệu có thể xác nhận tất cả trong 1 phút không.

Nhìn vào độ chính xác của việc phân loại mục tiêu, câu hỏi này trở nên nặng nề hơn. Theo Bloomberg, độ chính xác phân loại mục tiêu của Maven vào khoảng 60 phần trăm. Khi con người tự làm, độ chính xác được đưa tin là khoảng 84 phần trăm. Nghĩa là máy kém chính xác hơn con người. Cùng bản tin đó cho biết khi thời tiết xấu, có mây hoặc tầm nhìn mờ, độ chính xác ấy rơi xuống dưới 30 phần trăm. Chỉ ra mười mục tiêu thì bảy mục tiêu sai.

Máy chọn mục tiêu với độ chính xác 60 phần trăm rồi đưa ra trước mặt con người. Con người xem xét nó trong vòng 1 phút. Người xem xét xong đóng dấu phê chuẩn. Rồi chuyển sang mục tiêu tiếp theo. Một nghìn lần trong một ngày. Trong dòng chảy ấy, vai trò của con người là gì. Là người phán đoán, hay là người ký tên vào kết quả mà máy đưa ra.

Bộ trưởng Quốc phòng Hoa Kỳ Pete Hegseth đã đẩy dòng chảy ấy đi nhanh hơn. Ông được đưa tin là đã phát biểu theo ý rằng không có 'quy tắc giao chiến ngu ngốc'. Quy tắc giao chiến

() là những quy định xác định khi nào và bằng cách nào quân nhân được dùng vũ khí. Đó là cơ chế để bảo vệ dân thường và đặt giới hạn cho việc dùng vũ lực. Khoảnh khắc gọi nó là ngu ngốc, mọi thứ cản tốc độ đều trở thành vướng víu.

Ban đầu, Bộ Ngoại giao Iran khẳng định Khamenei 'an toàn'. Họ không thừa nhận cái chết của lãnh đạo. Việc ông qua đời được xác nhận chính thức vào ngày 1 tháng 3, thông qua Hội đồng An ninh Quốc gia Iran. Một tuần sau, ngày 8 tháng 3, Mojtaba Khamenei được chỉ định làm người kế nhiệm. Người con nối vị trí của người cha.

Việc Khamenei bị tiêu diệt giữa ban ngày, tức vào thời điểm ban ngày, cũng được đưa tin như một điều khác thường. Các vụ tiêu diệt mục tiêu thường diễn ra trong bóng tối. Vì bóng đêm che giấu cả con người lẫn máy móc. Nhưng lần này là ban ngày. Điều đó có nghĩa là thông tin đã chắc đến mức ấy, hoặc không còn lý do gì để chần chừ đến mức ấy.

Một cuộc tấn công loại bỏ ban lãnh đạo trong một lần như vậy được gọi trong thuật ngữ quân sự là chiến dịch chặt đầu. Nghĩa là đánh vào cái đầu. Mục tiêu là làm tê liệt cơ cấu chỉ huy của đối phương, khiến quân đội không nhận được mệnh lệnh. Để chiến dịch này thành công, hai điều phải khớp nhau. Phải biết chính xác lãnh đạo đang ở đâu vào lúc này, và lãnh đạo không được di chuyển trước khi thông tin ấy được chuyển thành hành động. Nếu dùng bàn tay con người để khớp hai việc này, mục tiêu thường đã rời chỗ trong lúc thông tin còn đang được gom lại. Vì thời gian giữa thông tin và hành động quá dài. Máy đã rút ngắn thời gian đó. Khoảng cách giữa khoảnh khắc xác định vị trí và khoảnh khắc bom rơi đã hẹp lại. Con số 60 giây là một tên gọi khác của khe hẹp ấy.

Có người sẽ nói chuyện này đã trở nên chính xác hơn về mặt nhắm bắn. Nói vậy là đúng. Trước đây, có khi phải ném bom cả một thành phố mới bắt được mục tiêu; nay có thể thu hẹp vào một tòa nhà rồi đánh. Đây cũng là công nghệ giảm thiệt hại cho dân thường. Nhưng độ tinh chuẩn và độ đúng không phải là một. Tinh chuẩn nghĩa là bắn trúng đúng nơi đã nhắm. Đúng nghĩa là nơi đã nhắm thật sự là nơi cần nhắm. Tên lửa ngày càng tinh chuẩn. Nhưng việc quyết định phải nhắm vào đâu, độ đúng của việc ấy, lại là vấn đề khác với tính năng của tên lửa. Đó là vấn đề của dữ liệu, của phán đoán, và của thời gian. Ngôi trường ở mục tiếp theo cho thấy sự khác nhau ấy.

Số người chết ngay từ đầu đã không thống nhất. Lúc đầu, Hội Trăng lưỡi liềm đỏ Iran thống kê khoảng 201 người thiệt mạng. Nhưng các thống kê tiếp sau lớn hơn nhiều. Bảng tổng hợp của Wikipedia ghi số người chết phía Iran là 3.468 người, số người bị thương là 26.500 người. 201 người và 3.468 người. Hai con số của cùng một sự kiện. Vào thời điểm viết những dòng này, khó có thể nói chắc bên nào đúng. Nó khác nhau tùy nguồn, và khác nhau tùy tiêu chuẩn thống kê. Nhưng riêng việc hai con số cách xa đến vậy đã nói lên chiến dịch này lan nhanh và rộng đến mức nào.

Có tin cho biết phía quân đội Mỹ cũng có người thiệt mạng. Chiến dịch này không phải là việc chỉ một bên chịu tổn thất.

60 giây. 1.000 mục tiêu. Độ chính xác 60 phần trăm. Đặt ba con số này cạnh nhau, ta thấy đường nét của một thời đại. Đó là thời đại mà quyết định nặng nề nhất của chiến tranh, quyết định giết ai, đã chuyển từ tốc độ của con người sang tốc độ của máy móc. Trong thời đại ấy, con người vẫn ngồi trước cò súng. Chỉ khác là trước mặt họ chất đống một nghìn tọa độ do máy đặt xuống, và thời gian dành cho con người là 1 phút cho mỗi tọa độ.

Mục tiếp theo của chương này cho thấy điều gì có thể xảy ra trong 1 phút ấy. Tại tọa độ của một ngôi trường.

## 2. Trường Minab: tọa độ cũ gọi đến 175 người

Hàng ghế đầu của lớp học, chỗ thứ hai bên cửa sổ. Sáng hôm ấy, trên bàn của đứa trẻ đó vẫn có một cuốn vở đang mở.

Minab, một thành phố cảng thuộc tỉnh Hormozgan ở miền nam Iran. Trường tiểu học Shajareh Tayyebeh. Sáng ngày 28 tháng 2 năm 2026, đồng hồ vừa qua 10 giờ 20 phút một chút. Đó là lúc tiết học thứ nhất kết thúc và tiết thứ hai sắp bắt đầu. Trẻ em đang ngồi tại chỗ. Các cô giáo đứng trước bảng.

Vào thời điểm ấy, khoảng từ 10 giờ 23 phút đến 10 giờ 45 phút theo giờ chuẩn Iran, một vũ khí được cho là tên lửa hành trình Tomahawk do Mỹ chế tạo đã rơi xuống tòa nhà của ngôi trường. Đây là nội dung được The New York Times và CNN đưa tin. Loại vũ khí được cho là thuộc dòng UGM-109, hệ Tomahawk.

Số người thiệt mạng được thống kê khác nhau. Các bản tin nêu con số từ khoảng 156 người đến hơn 175 người. Phần lớn là trẻ em từ bảy đến mười hai tuổi. Khoảng 73 bé trai, khoảng 47 bé gái. Và khoảng 26 cô giáo làm việc tại ngôi trường ấy đều thiệt mạng. Không một ai sống sót.

Trước những con số này, chúng ta phải thận trọng. Hiện khó thể khẳng định con số 156 hay 175 là chính xác. Việc thống kê khi ấy vẫn đang tiếp diễn, và mỗi nguồn đưa ra một con số khác nhau. Nhưng con số dao động không có nghĩa là những đứa trẻ có mặt ở đó cũng trở nên mơ hồ. Đứa trẻ bảy tuổi vẫn là đứa trẻ bảy tuổi. Đứa trẻ ngồi ở bàn thứ hai cạnh cửa sổ đã thật sự ngồi ở chỗ ấy.

Khi viết bằng số, 175 chỉ là một dòng. Nhưng nếu mở dòng ấy ra, nó trở thành 175 buổi sáng. 175 đứa trẻ đã đeo cặp rời nhà vào sáng hôm đó. Có em có lẽ ngủ dậy muộn nên chạy vội đến trường, có em vừa đi vừa lo bài tập hôm qua chưa giải xong. Có em có thể bước vào lớp với chuyện cãi nhau với bạn cùng bàn còn vướng trong lòng. Từ bảy đến mười hai tuổi. Đó là khoảng tuổi giữa lúc bắt đầu thay răng và lúc chiều cao lớn nhanh rõ rệt. Với những đứa trẻ ấy, những từ như Lực lượng Vệ binh Cách mạng hay cơ sở dữ liệu mục tiêu có lẽ là những từ cả đời

các em chưa từng nghe.

26 nữ giáo viên cũng đi làm vào sáng hôm đó. Trong số họ, có người có lẽ đã dạy ở ngôi trường ấy từ lâu, có người có lẽ là giáo viên trẻ mới được bổ nhiệm. Khi họ đứng trước bảng, chờ tiếng chuông tiết hai vang lên, họ không thể biết rằng mình đang hiện trên màn hình của một hệ thống quân sự nào đó bằng tọa độ của 10 năm trước. Tôi sẽ không miêu tả cảnh tượng ngày hôm ấy theo lối kích động. Làm vậy là tiêu dùng người chết thêm một lần nữa. Tôi chỉ muốn nói rõ một điều. Mỗi khi con số 175 xuất hiện trong cuốn sách này, đó không phải là thống kê, mà là 175 chiếc bàn học bỏ trống.

Chuyện như vậy đã xảy ra thế nào. Không phải vũ khí bay lệch. Tên lửa đã bay chính xác đến tọa độ. Vấn đề nằm ở chính tọa độ ấy.

The New York Times và CNN đưa tin nguyên nhân là các tọa độ cũ, không được cập nhật của Cơ quan Tình báo Quốc phòng Hoa Kỳ (DIA). Cơ quan Tình báo Quốc phòng là cơ quan xử lý thông tin quân sự. Trong cơ sở dữ liệu của cơ quan này, khu đất ở Minab được ghi là cơ sở quân sự. Cụ thể là căn cứ của Lực lượng Vệ binh Cách mạng.

Đã có thời, cách ghi đó là đúng. Cho đến năm 2016, ngôi trường ấy vẫn dùng chung khu đất với một căn cứ của Lực lượng Vệ binh Cách mạng. Cơ sở quân sự và trường học nằm trong cùng một hàng rào. Rồi khoảng năm 2016, hai nơi được tách ra. Quân đội rời đi, chỉ còn lại trường học.

Đó là thông tin của 10 năm trước. Suốt 10 năm, tọa độ ấy không được cập nhật. Trong cơ sở dữ liệu, khu đất ấy vẫn là căn cứ của Lực lượng Vệ binh Cách mạng. Với trí nhớ con người, 10 năm là quãng thời gian đủ để ký ức phai nhạt. Nhưng cơ sở dữ liệu của máy móc thì không phai. Nó chỉ không được cập nhật. Dữ liệu không được cập nhật vẫn còn lại, rõ ràng trong trạng thái sai. Và chính sự rõ ràng ấy nguy hiểm hơn. Vì thông tin mờ khiến con người nghi ngờ, còn tọa độ rõ khiến con người tin.

Trong dữ liệu không có thời gian. Hãy hình dung một bức ảnh vệ tinh. Trong đó có một tòa nhà được chụp lại, kèm theo tọa độ. Nhưng bản thân bức ảnh không ghi nó được chụp khi nào, cũng không ghi tòa nhà ấy trong khoảng thời gian sau đó đã biến thành nơi gì. Con người phải tự tìm hiểu riêng. Phải có ai đó nói rằng: "Thông tin tọa độ này là từ năm 2016, hãy kiểm tra lại." Trong một phút ấy, liệu có người nào ở đó để nói câu ấy không. Trong dòng chảy mỗi ngày có cả nghìn mục tiêu trôi qua, liệu có ai giữ lại một tọa độ và hỏi: "Thông tin này có từ bao giờ" không.

Máy móc không đặt câu hỏi đó. Với máy móc, tọa độ của năm 2016 và tọa độ của năm 2026 đều chỉ là những con số rõ ràng như nhau. Con số không có tuổi. Vì vậy, tọa độ cũ hiện lên trên màn hình mục tiêu với khuôn mặt của tọa độ mới. Rồi tên lửa rơi xuống đó.

Có một sự thật nặng nề hơn. Khi cuộc không kích diễn ra, website của ngôi trường ấy vẫn hoạt động. Trên bản đồ Internet, nơi đó cũng được đánh dấu là trường học. Đây là điểm CNN nêu ra. Nếu bất kỳ ai gõ tên ngôi trường ấy vào ô tìm kiếm, chỉ trong vài giây họ đã có thể biết đó là một trường tiểu học. Trang chủ của trường sẽ hiện ra, và ký hiệu trường học trên bản đồ cũng sẽ hiện ra.

Sự thật này về sau trở thành câu hỏi được đặt trực tiếp cho giám đốc điều hành của công ty tạo ra Claude. Ngôi trường ấy có website, tìm trên Google là thấy, vậy tại sao Claude, hoặc bất kỳ công nghệ nào họ đã dùng, lại không chỉ ra được điều đó. Ở phần cuối của chương này, chúng ta sẽ nói đến câu hỏi ấy và câu trả lời của ông ấy.

Điều cần nhìn thẳng bây giờ là cấu trúc trách nhiệm.

Người chế tạo tên lửa đã chế tạo tên lửa một cách chính xác. Tomahawk được thiết kế để bay đến tọa độ, và nó đã bay chính xác đến tọa độ. Hệ thống nhập tọa độ đã trung thành lấy tọa độ có trong cơ sở dữ liệu. Cơ quan xây dựng cơ sở dữ liệu đã nhập thông tin chính xác vào năm 2016. Người rà soát mục tiêu đã ký vào kết quả do hệ thống đưa ra. Người bóp cò đã làm theo mệnh lệnh.

Ai cũng làm đúng việc của mình ở đúng vị trí của mình. Nhưng 175 người đã chết.

Đây là một hình dạng của điều mà cuốn sách này gọi là 'chỗ trống'. Chỗ trách nhiệm biến mất. Không ai nói dối, không ai vi phạm quy tắc một cách rõ ràng, nhưng chỉ còn lại kết quả. Trách nhiệm không cập nhật dữ liệu thuộc về ai. Trách nhiệm không nghi ngờ một tọa độ đã 10 năm tuổi thuộc về ai. Trong dòng xử lý một nghìn mục tiêu trong vòng 1 phút, ai có trách nhiệm tách riêng một tọa độ ra và gõ vào ô tìm kiếm để kiểm tra.

Đại úy Tim Hawkins, người phát ngôn của Bộ Tư lệnh Trung tâm Hoa Kỳ (CENTCOM), đã trả lời báo chí về vụ việc này. Việc kênh phát ngôn chính thức của quân đội lên tiếng có nghĩa là chuyện này không bị chôn vùi như một tai nạn nhỏ.

Tổ chức nhân quyền quốc tế Human Rights Watch (HRW) vào tháng 4 năm 2026 đã phân tích liệu cuộc không kích này có cấu thành tội ác chiến tranh hay không. Cụm từ tội ác chiến tranh rất nặng. Trong luật quốc tế có nguyên tắc phải phân biệt giữa chiến binh và dân thường. Đó là nguyên tắc phân biệt. Cũng có nguyên tắc rằng lợi ích quân sự của cuộc tấn công phải lớn hơn thiệt hại gây ra cho dân thường. Đó là nguyên tắc tương xứng. Human Rights Watch đã đặt câu hỏi: cuộc tấn công nhầm một trường học thành cơ sở quân sự ấy có vi phạm những nguyên tắc này hay không.

Nguyên tắc phân biệt là nền móng lâu đời nhất của luật chiến tranh. Binh sĩ có thể tấn công binh sĩ, nhưng không được tấn công dân thường. Với trẻ em thì càng không cần nói thêm. Nhưng nguyên tắc này dựa trên một tiền đề. Phải phân biệt được ai là binh sĩ, ai là dân thường. Khi con người thực hiện việc phân biệt ấy, con người chịu trách nhiệm. Nếu phân biệt sai, người

đó phải trả lời. Nhưng nếu máy móc tham gia vào việc phân biệt, nếu máy móc nhìn vào dữ liệu đã 10 năm, còn con người chấp nhận kết quả ấy trong vòng 1 phút, thì trách nhiệm cho sự phân biệt sai sẽ đi về đâu. Nguyên tắc phân biệt vẫn còn đó, nhưng bàn tay thực sự làm việc phân biệt đã thay đổi. Nguyên tắc được viết hướng tới con người, nhưng công việc thì máy móc đã làm gần một nửa. Sự lệch pha ấy đã hiện ra trong lớp học ở Minab, dưới hình hài của 175 con người.

Ở đây, tôi muốn nói thêm một câu từ vị trí của một luật sư bào chữa. Lời biện minh rằng đó là nhầm lẫn không phải lúc nào cũng đứng vững trước pháp luật. Khi một người lái xe đâm vào người khác, chỉ nói "tôi không biết" thì không thể thoát trách nhiệm. Bởi người đó có nghĩa vụ phải biết. Việc dùng nguyên tọa độ đã có từ 10 năm trước có phải là lỗi bất cẩn hay không, và mức độ của lỗi ấy nặng đến đâu, đó là việc phải được làm rõ tại tòa. Nhưng có một điều rõ ràng. Hệ thống nhanh hơn không có nghĩa là nghĩa vụ phải biết cũng biến mất nhanh hơn.

Hãy nghĩ đến 26 nữ giáo viên của ngôi trường ấy. Sáng hôm đó họ đã đi làm. Họ đứng trước bảng. Có lẽ họ đã gọi tên từng đứa trẻ. Họ không có cách nào biết rằng mình vẫn còn nằm trong một cơ sở dữ liệu mục tiêu, dưới dạng những tọa độ từ 10 năm trước. Không ai có thể biết được. Đó là điều đáng sợ của thời đại này. Một người, trong lúc chính họ không hay biết, đã trở thành một tọa độ trong một cơ sở dữ liệu nào đó. Tọa độ ấy đã được cập nhật hay chưa sẽ quyết định số phận của người đó, nhưng người đó không thể biết, cũng không thể thay đổi.

Chúng ta quay lại với đứa trẻ ngồi ở chỗ thứ hai bên cửa sổ. Chỗ ấy giờ đã trống. Cuốn vở của em, cuốn vở đang mở ra ấy, cũng đã ở đó cùng em. Một chỗ ngồi trống của một đứa trẻ. Điều nhan đề cuốn sách này chỉ đến là nhiều chỗ trống. Chỗ trách nhiệm biến mất, và chỗ con người biến mất. Ở Minab, cả hai cùng hiện diện trong một lớp học.

## 3. Va chạm giữa Thung lũng Silicon và Bộ Quốc phòng: Claude và rủi ro chuỗi cung ứng

Một khoảng ba ngày nào đó trong tháng 2 năm 2026. Một công ty ở San Francisco nhận được tối hậu thư từ Washington. Hãy trả lời trong vòng ba ngày. Ba ngày ấy đang thu ngắn lại trên màn hình.

Những người trong phòng họp hẳn đã biết đồng thời hai điều. Nếu từ chối, công ty sẽ gặp nguy hiểm. Nếu chấp nhận, ranh giới do chính họ vạch ra sẽ sụp đổ. Không có con đường nào dễ chịu cả.

Bên gửi tối hậu thư là Bộ Quốc phòng Hoa Kỳ, cơ quan nay còn được gọi là 'Bộ Chiến tranh (Department of War)'. Bên nhận là Anthropic, công ty tạo ra Claude. Yêu cầu rất rõ ràng. Họ phải gỡ bỏ mọi giới hạn sử dụng của Claude.

Muốn hiểu cuộc va chạm này, trước hết phải nhìn vào một sự thật. Anthropic không phải là công ty giữ khoảng cách với quân đội. Ngược lại mới đúng.

Dario Amodei đã nói rõ điều này trong cuộc phỏng vấn với CBS. Ông nói Anthropic là công ty 'đi đầu nhất trong số mọi công ty AI trong việc làm việc với chính phủ Hoa Kỳ và quân đội Hoa Kỳ'. Nếu chuyển nguyên ý lời ông, thì Anthropic là công ty đầu tiên đưa mô hình của mình lên đám mây mật, là công ty đầu tiên tạo mô hình tùy chỉnh cho mục đích an ninh quốc gia, và các mô hình ấy đang được triển khai trong cộng đồng tình báo cũng như trong toàn quân, dùng cho những việc như hỗ trợ tác chiến mạng và chiến đấu.

Khi được hỏi vì sao làm như vậy, ông trả lời: 'Vì chúng tôi tin rằng chúng tôi phải bảo vệ đất nước mình'. Ông nói nước Mỹ phải được bảo vệ trước các quốc gia chuyên chế như Trung Quốc và Nga. Ông tự gọi mình là 'một người Mỹ yêu nước'.

Nhưng chính người Mỹ yêu nước ấy đã vạch ranh giới ở hai việc.

Việc thứ nhất là giám sát trong nước trên quy mô lớn. Trên CBS, ông giải thích như sau. Chính phủ mua dữ liệu do doanh nghiệp tư nhân thu thập, rồi dùng trí tuệ nhân tạo để phân tích tất cả cùng một lúc. Ông nói việc đó thật ra không bất hợp pháp. Chỉ là trong thời kỳ trước trí tuệ nhân tạo, nó không có ích mấy. Công nghệ chạy quá nhanh, còn luật chưa theo kịp. Vì vậy đó là một vùng nguy hiểm. Đó là ranh giới đầu tiên mà ông vạch ra.

Việc thứ hai là vũ khí hoàn toàn tự động. Ông phân biệt nó với các loại vũ khí bán tự động đang được dùng ở Ukraine. Theo ông, vũ khí hoàn toàn tự động là ý tưởng tạo ra 'vũ khí được phóng đi mà không có bất kỳ sự can thiệp nào của con người'. Đó là loại vũ khí không có con người nào trong chuỗi quyết định ai sẽ là mục tiêu, ai sẽ bị bắn. Ông lo ngại hai điều về loại vũ khí này. Một là độ tin cậy. Trí tuệ nhân tạo ngày nay chưa đủ đáng tin để tạo ra vũ khí hoàn toàn tự động. Ông nói trí tuệ nhân tạo có 'tính khó dự đoán căn bản', và về mặt kỹ thuật, điều đó vẫn chưa được giải quyết. Điều còn lại là giám sát. Nếu có một đội quân drone hoặc robot vận hành mà không có sự giám sát của con người, câu hỏi sẽ nảy sinh: ai chọn mục tiêu và ai chịu trách nhiệm.

Trong cuộc phỏng vấn với Bloomberg, ông ví nỗi lo ấy với một bộ phim. Đó là 'Dr. Strangelove'. Một thiết bị tận thế tự động phóng vũ khí hạt nhân khi nhận định rằng vũ khí hạt nhân đang bay tới. Ông hỏi ngược lại: có gì có thể sai được chứ? Chiến tranh có thể nổ ra theo cách hai nước hiểu lầm nhau rồi lao vào nhau; nếu công nghệ ấy không có sự giám sát đúng mức, những tai nạn như vậy càng dễ xảy ra.

Theo cách tính của Anthropic, 98 đến 99 phần trăm toàn bộ trường hợp sử dụng là có thể chấp nhận. Thứ họ muốn chặn lại là 1 phần trăm. Chỉ có hai việc: giám sát trong nước quy mô lớn và vũ khí hoàn toàn tự động.

Amodei giải thích bản chất của ranh giới này bằng một ví dụ. Đó là trong cuộc phỏng vấn với Bloomberg. Ông nói, câu hỏi này giống như hỏi rằng nếu kẻ địch phạm tội ác chiến tranh, thì chúng ta cũng phải phạm tội ác chiến tranh hay sao. Ông nói rõ rằng mình không khẳng định

vụ Minab là tội ác chiến tranh. Điều ông muốn nói là chuyện khác. Phải chiến thắng theo cách giữ được các giá trị của mình; không được lao vào cuộc đua xuống đáy. Dù thắng, nếu trong quá trình đó đánh mất điều mình từng muốn bảo vệ, thì đó không phải là chiến thắng. Trong cùng cuộc phỏng vấn, ông còn nói thế này. Nếu nền dân chủ làm những việc như vậy, thì việc nền dân chủ chiến thắng sẽ không còn đáng với cái giá phải trả.

Khi được hỏi điều gì có thể đi sai hướng, ông nêu ra hai điểm. Một là độ tin cậy. Đó là trường hợp máy móc nhắm nhầm người, bắn vào dân thường, và không thể hiện được phán đoán mà một người lính con người lẽ ra sẽ có. Bắn nhầm quân mình, đánh trúng dân thường, hoặc chỉ đánh nhầm mục tiêu. Ông nói mình không muốn bán thứ không thể tin được. Nghĩa là không muốn bán thứ có thể giết người của chính nước mình hoặc giết người vô tội. Điểm còn lại là sự giám sát. Với một người lính con người, luôn có một chuỗi trách nhiệm đi kèm. Chuỗi đó đứng trên giả định rằng người ấy sẽ dùng lẽ thường của mình. Nhưng nếu một người, hoặc vài người, điều khiển mười triệu drone, thì chuỗi trách nhiệm ấy sẽ móc vào đâu. Ông cho rằng khi quyền lực tập trung đến mức đó vào một nơi, trách nhiệm sẽ bị phân tán.

Bộ Quốc phòng không chấp nhận 1 phần trăm đó.

Theo lời Amodei, cuộc đàm phán diễn ra rất nhanh. Nó nằm trong một khung thời gian hẹp, chỉ ba ngày. Hai bên trao đổi vài bản câu chữ. Có thời điểm, Bộ Quốc phòng gửi một bản trông như chấp nhận điều kiện của Anthropic. Nhưng trong đó lại gắn thêm các điều kiện như "nếu Bộ Quốc phòng cho là phù hợp" hoặc "chừng nào còn phù hợp với pháp luật". Ông nói rốt cuộc đó là một bản trên thực tế không nhượng bộ gì cả.

Hãy đặt lại mốc thời gian. Bộ trưởng Hegseth yêu cầu gỡ bỏ mọi hạn chế sử dụng vào ngày 24 tháng 2. Đó là tối hậu thư với thời hạn ba ngày. Anthropic từ chối vào ngày 26 tháng 2. Tổng thống Hoa Kỳ phê chuẩn chiến dịch vào ngày 27 tháng 2, và chiến dịch bắt đầu vào ngày 28 tháng 2.

Khi đặt các ngày cạnh nhau, một điều hiện rõ. Ngày hôm sau khi Anthropic từ chối, chiến dịch được phê chuẩn; ngày kế tiếp, ngôi trường ở Minab sụp đổ.

Rồi đến ngày 4 tháng 3, chính quyền Trump chỉ định Anthropic là "rủi ro chuỗi cung ứng (supply chain risk)".

Amodei đã giải thích trên CBS rằng sự chỉ định này có nghĩa gì. Khi một công ty bị coi là rủi ro đối với chuỗi cung ứng, công ty đó phải bị loại khỏi mọi hệ thống của chính phủ. Ông nói rằng các sĩ quan quân đội tại hiện trường mà ông gặp đã nói như sau. Không có công cụ này, chiến dịch sẽ bị chậm 6 tháng, 12 tháng, có khi còn lâu hơn.

Amodei có lý do để tức giận về sự chỉ định này. Theo những gì ông biết, chưa từng có lần nào nhãn rủi ro chuỗi cung ứng được áp dụng cho một công ty Mỹ. Nhãn đó chỉ được dùng cho doanh nghiệp của các nước đối địch. Chẳng hạn Kaspersky Lab, công ty an ninh bị nghi có liên

hệ với chính phủ Nga, hoặc các nhà cung cấp bán dẫn Trung Quốc. Ông gọi việc Anthropic bị xếp chung với những công ty như vậy là "mang tính trả đũa và không phù hợp".

Tổng thống công khai chỉ trích "sự ích kỷ" của Anthropic, nói rằng điều đó "đặt mạng sống của người Mỹ vào nguy hiểm và làm quân đội của chúng ta lâm nguy". Trump còn gọi Anthropic là "công ty woke cánh tả". Bộ trưởng Hegseth đăng bài với ý rằng bất kỳ công ty nào có hợp đồng quân sự cũng không thể giao dịch với Anthropic. Amodei phản bác rằng đây là cách nói phóng đại so với những gì luật thực sự quy định. Theo ông, điều luật đặt ra một giới hạn hẹp hơn nhiều: công ty có hợp đồng quân sự không được dùng Anthropic như một phần của chính hợp đồng quân sự đó.

Có một điểm đáng để nhìn kỹ. Khi Anthropic từ chối, một công ty khác đã lấp vào chỗ trống. Amodei nói trên Bloomberg rằng OpenAI đã bước vào và ký hợp đồng mà Anthropic không muốn ký. Ranh giới mà một công ty vạch ra đã bị một công ty khác vượt qua. Ông thừa nhận rằng có giới hạn đối với những gì một hợp đồng có thể ngăn lại.

Dù biết giới hạn ấy, ông vẫn giải thích vì sao mình từ chối hợp đồng. Một hợp đồng không thể ngăn công ty khác vạch một ranh giới khác, nhưng việc ông làm là đặt ra một tiền lệ. Nghĩa là ông đưa ra trước công chúng sự phân biệt giữa các trường hợp sử dụng mà ông cho là chấp nhận được và các trường hợp khiến ông lo ngại. Và nhờ vậy, theo ông, trong Quốc hội đã xuất hiện một chuyển động lưỡng đảng. Đó là chuyển động muốn cấm bằng luật những việc mà họ lo ngại, muốn dựng lên các hàng rào bảo vệ. Ông nói mình không muốn gọi đây là một cuộc chiến. Không phải là cuộc chiến giữa doanh nghiệp tư nhân và chính phủ, mà là một cuộc thảo luận về việc chính phủ nên dùng trí tuệ nhân tạo như thế nào.

Ở đây có một nghịch lý cũ. Công ty làm ra vũ khí thường không quyết định vũ khí đó sẽ được dùng ở đâu. Boeing chế tạo máy bay quân sự, nhưng quân đội quyết định sẽ làm gì với chiếc máy bay ấy. Đó là phép so sánh mà người dẫn chương trình đặt ra với Amodei. Nhưng Amodei trả lời rằng trí tuệ nhân tạo thì khác. Có hai lý do. Một là công nghệ này quá mới. Máy bay là công nghệ đã lâu đời, nên các tướng lĩnh cũng hiểu khá rõ cách nó vận hành. Nhưng trí tuệ nhân tạo thay đổi quá nhanh, đến mức ngay cả người tạo ra nó cũng khó biết hết độ tin cậy và giới hạn của nó. Lý do còn lại là mô hình có một thứ có thể gọi là 'tính cách'. Có việc nó làm đáng tin, có việc thì không, và công ty tạo ra nó là bên hiểu rõ nhất điều đó.

Chính ở đoạn này, mâu thuẫn của cuốn sách hiện ra rõ nhất.

Công ty bán công nghệ đã vạch ranh giới trước cách sử dụng nguy hiểm nhất của công nghệ ấy. Khi một công ty vạch ranh giới, công ty khác vượt qua ranh giới đó và giành hợp đồng. Vì vậy việc lựa chọn mục tiêu không dừng lại. Công ty vạch ranh giới bị chính phủ trả đũa, còn công ty vượt qua ranh giới thì có hợp đồng. Trong cấu trúc ấy, lương tâm của một công ty có thể ngăn được điều gì. Gần như không ngăn được gì cả. Nó chỉ ném ra trước thế giới một câu hỏi.

Vậy rốt cuộc Claude có được dùng ở ngôi trường của Minab hay không?

Các người dẫn chương trình của CBS và Bloomberg đã hỏi thẳng Amodei. Nếu đúng như Bloomberg đưa tin, Claude được đưa vào hệ thống Maven Smart System của Palantir và được dùng để chọn mục tiêu trong chiến dịch Iran, thì Claude có góp phần vào cuộc không kích hồi tháng 2 đánh trúng trường học và giết chết hơn 150 trẻ em hay không?

Amodei trước hết nêu rõ một điều. Theo những gì ông biết, Anthropic không làm việc với ICE hay các cơ quan kiểm soát biên giới, và cũng không làm việc ở Gaza, ông nói trên Bloomberg. Ý ông là công ty đặt phạm vi tham gia của mình khá hẹp. Nhưng trước câu hỏi trực tiếp về chiến dịch Iran, câu trả lời của ông đã khác.

Câu trả lời của ông là ông không biết.

Ông nói như sau. Chúng tôi không thể biết chính xác các mô hình này đã được dùng như thế nào. Chúng tôi không có quyền truy cập như vậy. Những sai lầm kiểu này trong chiến tranh thật sự, thật sự khủng khiếp. Trong cùng cuộc phỏng vấn, ông nói chính vụ việc này cho thấy vì sao họ phải vạch ra ranh giới. Nghĩa là họ đã cố hạn chế cách sử dụng mô hình, dù phải đặt cả tương lai công ty vào đó. Rồi ông nhấn mạnh một điểm. Ngay cả trong vụ Minab, cách dùng đó cũng không vi phạm lằn ranh đỏ của họ. Claude có thể đã hỗ trợ, nhưng quyết định cuối cùng là do con người đưa ra. Con người là bên bóp cò sau cùng.

Người dẫn chương trình không lùi bước. Ngôi trường đó có website, có thể tìm thấy bằng Google Search, vậy Claude, hoặc bất kỳ AI nào mà họ dùng, lẽ ra phải phát hiện ra điều đó chứ, người dẫn hỏi. Chẳng phải chuyện này cho thấy mặt đáng sợ hơn của việc dùng công nghệ làm lối tắt trong chiến tranh hay sao?

Amodei lùi lại một bước, nói rằng câu hỏi này đòi hỏi ông phải dựa vào kiến thức mật mà ông không có. Ông nhắc lại rằng ông không biết Claude, hay bất kỳ AI nào khác, đã đóng vai trò gì. Tuy vậy, ông nói ông tin rằng nguyên tắc do họ đặt ra, nguyên tắc con người đưa ra quyết định cuối cùng, đã được giữ. Nếu vụ việc này không phải là ví dụ cho thấy vì sao nguyên tắc ấy quan trọng, thì còn điều gì là ví dụ nữa, ông hỏi ngược lại.

Khi được hỏi nếu có thể nói một câu với Tổng thống thì ông sẽ nói gì, Amodei trả lời như sau. Chúng tôi là những người Mỹ yêu nước. Mọi việc chúng tôi làm đều là vì đất nước này. Chúng tôi đi đầu làm việc với quân đội cũng vì chúng tôi tin vào đất nước này. Và chúng tôi vạch ra hai lằn ranh đỏ vì tin rằng vượt qua những lằn ranh ấy là trái với các giá trị của nước Mỹ. Khi chính phủ đe dọa bằng sự can thiệp chưa từng có, gồm việc chỉ định rủi ro chuỗi cung ứng và Đạo luật Sản xuất Quốc phòng, chúng tôi chỉ thực hiện quyền tự do ngôn luận được Hiến pháp bảo đảm để lên tiếng. Ông nói, không đồng ý với chính phủ chính là điều Mỹ nhất.

Trong câu nói ấy có cốt lõi của va chạm này. Ông làm việc với quân đội để bảo vệ đất nước, và cũng vì lý do đó, ông từ chối yêu cầu của quân đội. Cùng một lòng yêu nước đã tách ra theo hai

hướng. Chế tạo vũ khí để ngăn kẻ địch cũng là bảo vệ đất nước; đặt giới hạn cho thứ vũ khí ấy để giữ giá trị của đất nước cũng là bảo vệ đất nước. Chính phủ chỉ xem điều thứ nhất là yêu nước, còn ông xem điều thứ hai cũng là yêu nước. Sự lệch pha ấy, chỉ trong ba ngày, đã đặt một công ty đứng cạnh danh sách doanh nghiệp của nước đối địch.

Ta nên đọc câu trả lời này như thế nào.

Ở một mặt, ông đã thẳng thắn. Ông nói mình không biết điều mình không biết. Có thể họ thật sự không có quyền nhìn vào Claude đã được dùng như thế nào, trên tọa độ nào. Sau khi bán một mô hình, ngay cả người tạo ra nó cũng không biết hết mô hình ấy sẽ đi đâu.

Nhưng ở mặt khác, chính sự thẳng thắn ấy phơi ra khoảng trống của thời đại này. Nếu ngay cả công ty tạo ra mô hình cũng không biết mô hình ấy đã được dùng vào việc gì, thì trách nhiệm nằm ở đâu. Người tạo ra nói rằng mình không biết. Công ty đưa nó vào hệ thống nói rằng họ chỉ giúp chọn mục tiêu. Hệ thống nhập tọa độ nói rằng nó chỉ làm theo cơ sở dữ liệu. Cơ sở dữ liệu chỉ lưu giữ trung thực thông tin từ 10 năm trước. Người ra quyết định cuối cùng chỉ ký vào kết quả mà hệ thống đưa ra. Và những kết quả ấy đang đổ ra với tốc độ một nghìn mục tiêu mỗi 24 giờ.

Câu nói rằng con người đã ra quyết định cuối cùng nghe có vẻ khiến ta yên tâm. Không phải máy móc, mà là con người bóp cò. Nhưng nếu người đó chỉ có 1 phút cho mỗi mục tiêu, và nếu độ chính xác của mục tiêu do máy đưa ra chỉ là 60 phần trăm, thì cái gọi là 'quyết định cuối cùng của con người' ấy là một tấm khiên dày đến mức nào. Đó có phải là vị trí của người chịu trách nhiệm không, hay chỉ là nơi trách nhiệm trôi qua rồi tạm dừng lại trong chốc lát.

Amodei còn vẽ ra một cảnh khác trên Bloomberg. Nếu đó là một thế giới cho phép vũ khí hoàn toàn tự chủ, nếu đó là một thế giới nơi trí tuệ nhân tạo ra quyết định còn con người thậm chí không nhìn thấy quyết định ấy, thì một việc như Minav sẽ trở thành thế nào. Điều ông sợ là thế giới đó. Theo ông, hiện tại, khi con người còn ít nhất nhìn lại ở bước cuối, vẫn tốt hơn thế giới nơi con người bị loại ra hoàn toàn.

Ông cũng ví chuyện này với việc chính phủ dùng Microsoft Excel. Chính phủ dùng Excel rất nhiều. Nhưng công ty làm ra Excel không thể quyết định rằng 'chiến dịch quân sự này được dùng Excel, còn chiến dịch kia thì không được dùng'. Lập luận là người tạo ra công cụ không thể quyết định mọi cách dùng của công cụ ấy. Nghe qua thì có lý. Không thể bắt công ty làm ra chương trình bảng tính chịu trách nhiệm về chiến tranh.

Nhưng Claude không phải Excel. Excel chỉ cộng trừ các con số do con người nhập vào; nó không chọn giúp nên đánh ai. Bloomberg đưa tin rằng Maven Smart System có Claude bên trong đã chọn mục tiêu và đưa ra trước mặt con người. Giữa một công cụ cộng trừ và một công cụ chọn con người để đưa ra có một con sông khó vượt qua. Chính Amodei chắc hẳn biết con sông ấy tồn tại, nên ông mới vạch ra hai lằn ranh đỏ. Nếu là Excel, sẽ không cần lằn ranh đỏ.

Lời ông ấy có thể đúng. Một thế giới vắng con người sẽ nguy hiểm hơn. Nhưng lớp học ở Minab đã xảy ra trong một thế giới nơi con người chưa hề biến mất. Trong một thế giới vẫn còn nguyên tắc rằng con người đưa ra quyết định cuối cùng, 175 người đã chết. Con người đã có mặt ở đó. Trong một phút.

Khi khép lại chương này, xin để lại một điều. Bloomberg cho biết một quan chức Mỹ nói rằng nhờ sự hỗ trợ của trí tuệ nhân tạo, số mục tiêu mà quân đội Mỹ có thể tấn công trong một ngày đã tăng từ 1.000 lên 5.000. Người dẫn chương trình hỏi Amodei: vậy có nghĩa là Claude đang giúp giết nhiều người hơn, nhanh hơn; điều đó không khiến ông thấy bất an sao.

Amodei trả lời rằng ông ủng hộ năng lực giúp nước Mỹ trở nên hiệu quả hơn về mặt quân sự. Ông nói năng lực ấy không gây ra chiến tranh, mà ngăn chặn chiến tranh. Ông một lần nữa nói rằng mình là người yêu nước. Đồng thời, ông nói rõ một điều. Người cung cấp công nghệ là họ, nhưng việc quyết định có thể dùng công nghệ ấy cho chiến dịch quân sự nào, và chiến dịch nào thì không, không phải phần việc của họ. Chính sách phải được đặt trong tay những người có quyền ra quyết định quân sự.

Người tạo ra công nghệ không quyết định chính sách. Người quyết định chính sách không tạo ra công nghệ. Hệ thống nhập tọa độ không chịu trách nhiệm. Cơ sở dữ liệu không nghi ngờ. Và ở đâu đó giữa những khoảng cách ấy, 175 chỗ trống đã xuất hiện.

Hãy gom lại những gì chúng ta đã thấy trong chương này. Một hệ thống có độ chính xác phân loại mục tiêu 60 phần trăm đã chọn mục tiêu trong tình trạng kém chính xác hơn con người. Khi thời tiết xấu, độ chính xác của hệ thống ấy rơi xuống dưới 30 phần trăm. Dù vậy, hệ thống ấy vẫn được đưa ra chiến trường, đặt ban lãnh đạo của một quốc gia và lớp học của một ngôi trường lên cùng một màn hình như những mục tiêu. Một công nghệ chưa hoàn thiện đang được thử nghiệm không phải trong phòng thí nghiệm, mà trong một cuộc chiến thật.

Cuốn sách này gọi tình trạng ấy là 'lethal beta'. Dịch sang tiếng Việt, có thể hiểu là bản thử nghiệm chết người. Khi làm phần mềm, trước khi phát hành chính thức, người ta đưa ra bản beta để người dùng thử. Có lỗi thì sửa. Beta là dấu hiệu cho thấy sản phẩm chưa hoàn thiện. Nhưng nếu bản beta của một hệ thống chọn mục tiêu có lỗi, thứ được sửa là phiên bản sau. Người trả giá cho lỗi ấy là người đang ở trên tọa độ đó. Ngôi trường ở Minab là lỗi của một bản beta. Lỗi mang tên độ chính xác 60 phần trăm, lỗi mang tên tọa độ 10 năm tuổi, lỗi mang tên một phút thời gian. Chỉ có điều, lỗi này không thể được hoàn tác bằng một bản vá. Những chiếc bàn trống sẽ không được lấp đầy ở phiên bản sau.

Có lẽ Amodei đã cố làm điều đúng. Ông đã vạch ra một ranh giới, và vì ranh giới ấy, công ty của ông bị chính phủ trả đũa. Ông thẳng thắn, và điều gì không biết thì nói là không biết. Đồng thời, ông đã bán công nghệ, rồi nói rằng việc công nghệ ấy được dùng vào đâu không phải phần việc của mình. Hai điều ấy cùng tồn tại trong một con người. Chúng ta có thể vẽ ông như anh hùng, cũng có thể vẽ ông như kẻ ác. Nhưng cả hai bức tranh đều sai. Ông là người đã nhìn

gần nhất vào những chỗ trống, và cũng là người không thể lấp đầy những chỗ trống ấy. Có lẽ đó là chỗ mà lương tâm của một người không thể tự mình lấp nổi.

Đó đáng lẽ là chỗ của ai. Chỗ mà trách nhiệm phải ngồi vào.

Trong chương tiếp theo về Iran, chúng ta sẽ thấy khoảng trống này vượt qua biên giới của một quốc gia và chuyển sang một chiến trường khác như thế nào. Đó là Venezuela. Ở đó, cùng một hệ thống, với cùng tốc độ, tiếp tục để lại cùng một câu hỏi.

# Phần 2 Vượt qua ranh giới kiểm soát Chương 4 Venezuela: Chiến dịch chặt đầu có AI can dự

## 1. Chiến dịch Quyết tâm Tuyệt đối: đêm bắt giữ Maduro

Toàn bộ ánh đèn ở Caracas vụt tắt cùng một lúc.

Đó là đêm 3 tháng 1 năm 2026. Trực thăng hạ xuống phía trên thành phố. Chúng bay ở độ cao khoảng 100 feet tính từ mái nhà, tức hơn 30 mét một chút. Trước hết là tiếng cánh quạt xé ngang thành phố ở tầm thấp, rồi mất điện xảy ra. Đèn giao thông tắt ngấm. Cửa sổ các tòa nhà công quyền chuyển sang màu đen. Ở Fort Tiuna, khu căn cứ quân sự ở ngoại ô Caracas nơi đặt Bộ Quốc phòng Venezuela, radar phòng không cũng biến mất khỏi màn hình đúng vào khoảnh khắc ấy.

Các bản tin cho biết ngay trước chiến dịch, một cuộc tấn công mạng đã làm tê liệt lưới điện và hệ thống phòng không của Caracas. Máy bay tiến vào bầu trời một thành phố đã mất điện, trong khi radar vốn phải phát hiện máy bay ấy lại nhắm mắt trước. Về sau, Chủ tịch Hội đồng Tham mưu trưởng Liên quân Mỹ Dan Caine dùng cụm từ "bất ngờ hoàn toàn" để nói về chiến dịch này.

Bất ngờ nghĩa là đối phương không biết. Muốn đối phương không biết, phải che mắt họ trước. Đêm ấy, đôi mắt của Caracas khép lại cùng lúc điện bị cắt. Thành phố không nhìn thấy thứ đang tiến đến.

Hãy hình dung Fort Tiuna trong chốc lát. Đây là trái tim của quân đội Venezuela bên trong Caracas. Ở đó có Bộ Quốc phòng, có các sở chỉ huy, có bệnh viện quân đội và doanh trại tập trung. Bản cáo trạng ghi rằng đây cũng là nơi những người thân cận của Maduro đưa thủ lĩnh Lực lượng Vũ trang Cách mạng Colombia đến gặp. Một vị trí nơi sức mạnh quân sự của một quốc gia được tập trung dày đặc, một vị trí bàn tay bên ngoài khó chạm tới. Nơi ấy đã bị xuyên thủng vào đêm 3 tháng 1. Việc nơi kiên cố nhất của đất nước bị xuyên thủng nói lên bản chất của chiến dịch này. Không phải một con phố bình thường ở Caracas, cũng không phải rìa ngoài của quân đội, mà lực lượng đặc nhiệm nước ngoài đã tiến vào tận phần sâu bên trong ấy. Thứ mở đường vào đó là thông tin được hợp nhất với cuộc tấn công mạng.

Đêm đó, 150 máy bay Mỹ cất cánh. Chúng xuất phát từ 20 căn cứ.

150 chiếc.

Hai mươi địa điểm.

100 feet trên mặt nước.

Nếu dồn những con số này vào cùng một nhịp thở, người đọc chỉ còn lại ấn tượng rằng đó là một chiến dịch lớn. Nhưng khi tách ra từng con số, câu chuyện khác hẳn. Muốn 150 máy bay cất

cánh cùng một ngày, cùng một thời điểm, hai mươi căn cứ phải vận hành theo một thứ tự đã được sắp sẵn. Căn cứ nào cất cánh trước, đi vào tuyến bay nào, hợp lưu tại điểm nào, chỉ cần lệch một chút thì 150 máy bay có thể va vào nhau. Việc sắp thứ tự ấy và ngăn sai lệch không thể được xử lý bằng một tờ giấy trong đầu con người.

Theo NBC, Cơ quan Tình báo Trung ương Mỹ (CIA) từ tháng 8 năm trước đã đưa các nhóm nhỏ bí mật vào Venezuela để theo dõi lịch trình di chuyển của Maduro. Nói cách khác, trong năm tháng, họ quan sát một người ngủ ở đâu, vào lúc nào, và di chuyển khi nào. Ông ta ở lại dinh thự nào trong mấy ngày, đi theo lộ trình nào, bên cạnh có ai, tất cả được tích lại thành hồ sơ của năm tháng.

Và lúc 10 giờ 46 phút tối ngày 3 tháng 1, theo giờ miền Đông nước Mỹ, cùng bản tin đó viết rằng Tổng thống Trump đã ra lệnh phóng cuối cùng vào đúng thời điểm ấy.

10 giờ 46 phút.

Đằng sau thời điểm tính bằng từng phút khi mệnh lệnh của một người được đưa ra là năm tháng theo dõi và sự chờ sẵn của hai mươi căn cứ. Mệnh lệnh thì ngắn. Sự chuẩn bị khiến mệnh lệnh ấy trở nên khả thi thì dài.

Cái khung lớn do Bộ Tư lệnh miền Nam Mỹ (SOUTHCOM) và lực lượng đặc nhiệm liên quân dựng sẵn mang tên 'Operation Southern Spear'. Bên trong đó, một chiến dịch riêng nhằm bắt một mình Maduro được vận hành, và cái tên xuất hiện trong bản tin là 'Operation Absolute Resolve'. Đó là một chiến dịch mang tên quyết tâm. Quyết tâm nằm trong lòng người, nhưng thứ đưa nó ra thực địa lại là máy bay từ hai mươi căn cứ và thông tin tụ lại trên một màn hình.

Tôi đã đưa cụm từ 'chiến dịch chặt đầu' vào nhan đề chương này. Trong thuật ngữ quân sự, chiến dịch chặt đầu chỉ cuộc tấn công nhằm loại bỏ ngay phần đầu của đối phương, tức ban lãnh đạo, để làm tê liệt toàn bộ tổ chức. Ý tưởng nằm ở chỗ không đối đầu từng phần với thân mình, mà chỉ đánh vào cái đầu. Lập luận ở đây là nếu bắt được một mình Maduro, thì toàn bộ 'cartel' mà ông ta được cho là đứng trên đỉnh sẽ sụp đổ. Nhưng như đã thấy ở phần trước, nếu thực thể của 'cartel' ấy mờ nhạt, thì thân mình sẽ sụp đổ khi cái đầu bị đánh cũng trở nên mờ nhạt. Chiến dịch chặt đầu chỉ đứng vững khi đầu và thân mình được xác định rõ. Nếu chỉ có cái đầu hiện rõ, còn thân mình chỉ là một biệt danh, thì rốt cuộc chiến dịch ấy đã đánh trúng cái gì?

Con số bay thấp 100 feet cũng không phải tự nhiên mà có. Radar khó bắt được vật thể bay thấp, vì chúng bị địa hình che khuất. Việc 150 máy bay tiến vào ở độ cao 30 mét trên mặt nước có nghĩa đó là chuyến bay nhằm tránh bị phát hiện. Nếu cộng thêm việc mạng lưới phòng không đã bị làm tê liệt trước bằng tấn công mạng, Caracas đã bị che mắt hai lớp. Một lớp là giới hạn của radar, không nhìn thấy máy bay bay thấp. Lớp còn lại là những radar đã bị tắt hẳn. Thiết kế hai lớp này cùng lúc chính là tác chiến tích hợp. Nếu độ cao của máy bay và thời điểm tấn công mạng lệch nhau, yếu tố bất ngờ sẽ vỡ. Việc khớp hai thứ ấy trên cùng một màn hình,

lại là phần việc của tấm kính đó.

Delta Force tiến vào dinh thự ở Forte Tiuna. Theo các bản tin, Maduro đã không kịp tới phòng an toàn bằng thép. Câu chuyện là ông ta thiếu vài giây ngay trước cánh cửa thép. Người đứng đầu một quốc gia đã chạy trong chính dinh thự của mình về phía cửa phòng an toàn, rồi bị bắt trước khi tay chạm tới cánh cửa ấy. Ông ta bị bắt cùng vợ, Cilia Flores.

Tạm dừng ở đây một chút. Vài giây trước cánh cửa thép. Cảnh tượng trong đó vài giây ấy chia đôi số phận một con người không phải là sự thật đã được kiểm chứng, mà là tình tiết do các bản tin thuật lại. Ông ta chính xác bị bắt ở đâu, bằng cách nào, khoảng cách tới cửa phòng an toàn là mấy bước, tất cả đều dựa vào lời giải thích của bên tiến hành chiến dịch và bên đưa tin. Tôi sẽ không khẳng định chắc. Tuy vậy, bức tranh mà các bản tin vẽ ra thì rõ ràng. Trong năm tháng, đường đi nước bước của một người đã bị theo dõi, và ở điểm cuối của quá trình theo dõi ấy, ông ta đã không chạm được tới cánh cửa thép.

Hai người bị áp giải tới căn cứ Không quân Vệ binh Quốc gia Stewart ở bang New York. Sau đó họ bị truy tố với cáo buộc khủng bố ma túy. Người cai trị trên thực tế của một quốc gia bị lực lượng đặc nhiệm nước ngoài bắt ngay trong khu quân sự của nước mình giữa đêm, rồi được đưa xuống một căn cứ quân sự ở châu lục khác trước khi trời sáng.

Cần nói qua Nicolás Maduro là người như thế nào. Ông xuất thân là tài xế xe buýt, tham gia hoạt động công đoàn rồi bước vào chính trị. Dưới thời Hugo Chávez, ông từng giữ chức ngoại trưởng và phó tổng thống; khi Chávez qua đời năm 2013, ông kế nhiệm chức tổng thống. Cả cuộc bầu cử năm 2018 lẫn năm 2024 đều bị cộng đồng quốc tế phê phán là gian lận, và hơn 50 quốc gia, trong đó có Mỹ, không công nhận ông là nguyên thủ hợp pháp. Cuốn sách này không nhằm bênh vực sự cai trị của ông. Nhưng nếu không ghi lại ông là ai, và những đánh giá quanh ông phân rẽ ra sao, ta sẽ nhìn chiến dịch đêm 3 tháng 1 chỉ bằng một con mắt. Với nhiều người Venezuela, ông là gương mặt của áp bức; đồng thời với Mỹ, ông là mục tiêu mà họ đã muốn bắt từ lâu. Hai điều này có thể cùng đúng.

Cần nhìn lại chi tiết theo dõi trong năm tháng. Nếu CIA đã vận hành đội xâm nhập từ tháng 8 năm 2025, thì trong năm tháng ấy, đời sống hằng ngày của Maduro đã được tích lại thành từng dòng dữ liệu. Ông ta ngủ ở tòa nhà nào, đội bảo vệ đổi ca vào giờ nào, đèn phòng nào tắt lúc nào. Những ghi chép như thế được thu thập bằng mắt người. Nhưng việc gom các ghi chép ấy lại, đọc ra mẫu hình, rồi chọn đêm nào có khe hở lớn nhất, nay không còn chỉ dựa vào đầu óc con người. Khi chồng năm tháng di chuyển lên cùng một màn hình, quy luật trong đời sống của một người sẽ hiện ra. Khi quy luật hiện ra, khe hở cũng hiện ra. Đêm 3 tháng 1 có lẽ chính là khe hở ấy.

Ngay từ việc gọi cảnh tượng này là gì cũng đã khó. Chính phủ Mỹ coi đây là hành động thực thi pháp luật đối với một nghi phạm khủng bố ma túy. Phía Maduro và chính phủ Venezuela coi đây là việc quân đội nước ngoài bắt cóc người đứng đầu một quốc gia có chủ quyền. Hai cách

định nghĩa va thẳng vào nhau. Một bên gọi là bắt giữ, bên kia gọi là xâm lược. Cuốn sách này không phải nơi phân xử cách định nghĩa nào đúng, cách nào sai. Tôi chỉ ghi rõ một sự thật: hai cách định nghĩa ấy đang chỉ cùng một hành động trong cùng một đêm.

Dòng chảy mà cuốn sách này lần theo là quá trình việc chọn mục tiêu chuyển từ tốc độ của con người sang tốc độ của máy móc. Ở Gaza, một hệ thống tên Lavender đưa con người vào danh sách, rồi sĩ quan tình báo đóng dấu chỉ trong 20 giây. Ở Ukraine, chuỗi giết chóc khép lại trong 18 phút. Ở Iran, 60 giây đã chia đôi số phận của một con người.

Đêm ngày 3 tháng 1 ở Venezuela là một hình dạng khác của chuỗi ấy.

Ở đây, tên lửa không rơi xuống đầu một người. Mục tiêu không bị giết mà bị bắt. Vì vậy chiến dịch này khác giọng với các chương trước. Điều chúng ta thấy ở Gaza và Iran là tốc độ con người bị giết. Điều thấy ở Venezuela là độ chính xác trong việc bắt người. Nhưng cấu trúc chống đỡ cho độ chính xác ấy thì giống nhau. Gom thông tin rải rác vào một màn hình, sắp xếp ai ở đâu và cần nhìn vào điều gì. Năm tháng truy đuổi, các đợt xuất kích từ hai mươi căn cứ, và cuộc tấn công mạng cắt lưới điện đều di chuyển trên một bảng hình ảnh đã được hợp nhất.

Chính bảng hình ảnh ấy là nhân vật thật sự của chương này.

Việc bắt thay vì giết có thể khiến chiến dịch này trông sạch sẽ. Việc tin tức về thương vong không nổi lên lớn như các chương trước cũng vậy. Nhưng trông sạch sẽ và trách nhiệm rõ ràng là hai chuyện khác nhau. Theo dõi một người suốt năm tháng, cắt lưới điện của đất nước người đó, cho máy bay cất cánh từ hai mươi căn cứ rồi lôi người đó ra giữa đêm, tất cả đều chứa vô số phán đoán. Càng ngày càng khó phân định đến bước nào phán đoán ấy còn là của con người, và từ bước nào nó đã thuộc về công cụ trên màn hình.

Cần ghi thêm một điều nữa. Lưới điện Caracas bị cắt không có nghĩa chỉ dinh tổng thống của Maduro tối đi. Cả thành phố tối đi. Trên những con đường nơi đèn tín hiệu tắt lịm có bệnh viện, có bệnh nhân sống nhờ máy thở, có những người thức giấc giữa đêm mà không biết chuyện gì xảy ra nên hoảng sợ. Tắt đèn của cả một thành phố để bắt một người. Hàng triệu người mắc kẹt ở giữa không phải mục tiêu của chiến dịch, nhưng chịu tác động từ chiến dịch. Đêm ấy của họ không được ghi lại như mục tiêu trong bất kỳ hồ sơ nào. Họ cũng không nằm trong thống kê thương vong. Nhưng sự thật rằng một thành phố bị ném vào bóng tối mà không hiểu vì sao là cái bóng luôn bám quanh những chiến dịch tự nhận là chính xác. Kích thước của cái bóng này đến nay chưa cơ quan truyền thông nào thống kê rõ ràng. Nó được để lại như một phạm vi chưa được xác nhận.

Tên của bảng hình ảnh ấy sẽ sớm xuất hiện. Trước hết cần nhìn vào lý do được đưa ra để bắt Maduro.

## 2. Sự sụp đổ của lý do: hư cấu mang tên Cartel Mặt Trời

Danh nghĩa bắt đầu từ một tờ giấy.

Đó là bản cáo trạng được nộp lên Tòa án Quận Liên bang khu vực Nam New York của Hoa Kỳ. Bản đầu tiên được nộp vào tháng 3 năm 2020. Gần sáu năm trước khi Maduro bị bắt, căn cứ pháp lý để bắt ông ta đã được chuẩn bị trước trên giấy. Trên bìa ghi: Hợp chúng quốc Hoa Kỳ kiện Nicolás Maduro Moros cùng sáu người khác. Đại bồi thẩm đoàn viết như sau.

"Trong hơn 25 năm, các nhà lãnh đạo Venezuela đã lạm dụng vị trí được công chúng tin cậy, làm tha hóa những định chế từng chính đáng, và đưa hàng tấn cocaine vào Hoa Kỳ."

Bản cáo trạng đặt Maduro ở tuyến đầu của sự tha hóa đó. Văn bản viết rằng, khi còn là nghị sĩ, ông ta đã vận chuyển cocaine dưới sự bảo vệ của các cơ quan thực thi pháp luật Venezuela. Khi làm ngoại trưởng, ông ta cấp hộ chiếu ngoại giao cho những kẻ buôn ma túy, và cung cấp vỏ bọc ngoại giao cho các chuyến bay đưa tiền ma túy từ Mexico trở lại Venezuela. Rồi khi trở thành tổng thống, và nay là nhà cai trị trên thực tế, ông ta đã để tình trạng tha hóa do cocaine thúc đẩy phát triển vì lợi ích của chính mình, của thân tín và gia đình mình.

Bản cáo trạng liệt kê liên tiếp những cảnh cụ thể. Có đoạn nói rằng năm 2006, hơn 5,5 tấn cocaine được chất lên một máy bay phản lực DC-9 để đưa sang Mexico, và số hàng đó bị thu giữ tại sân bay Ciudad del Carmen ở Campeche, Mexico. Văn bản còn mô tả chiếc máy bay ấy cất cánh sau khi nhận cocaine chở trên năm xe van tại nhà chứa máy bay dành cho chuyên cơ tổng thống ở sân bay Maiquetía. Có đoạn nói rằng vợ của Maduro, Flores, trong một cuộc họp năm 2007, đã nhận hàng trăm nghìn đô la tiền hối lộ và sắp xếp cuộc gặp giữa một trùm buôn ma túy lớn với người đứng đầu cơ quan chống ma túy. Cũng có đoạn nói rằng con trai của Maduro, Nicolás Ernesto Maduro Guerra, biệt danh 'Nicolasito' hoặc 'Hoàng tử', đã gặp các đại diện của Lực lượng Vũ trang Cách mạng Colombia (FARC) tại Medellín vào năm 2020 để bàn cách vận chuyển cocaine và vũ khí trong sáu năm tiếp theo, tức đến năm 2026.

Năm 2026. Năm được ghi trong bản cáo trạng ấy thật trùng hợp. Trên giấy, kế hoạch kéo dài đến năm 2026 đã được nhắc tới, và trên thực tế, chiến dịch đã diễn ra chỉ trong ba ngày đầu tiên của năm 2026. Đó là ngẫu nhiên hay có chủ ý, tôi sẽ không khẳng định.

Câu chữ trong bản cáo trạng bộc lộ nguyên vẹn niềm tin chắc của bên công tố. Về Maduro, đại bồi thẩm đoàn viết như sau.

"Nicolás Maduro Moros, bị cáo, đã đứng ở tuyến đầu của sự tha hóa đó, cùng các đồng phạm của mình sử dụng quyền lực giành được bất hợp pháp và những định chế mà ông ta đã làm mục ruỗng để vận chuyển hàng nghìn tấn cocaine vào Hoa Kỳ."

Các công tố viên đã chọn từ "ăn mòn". Ý của họ là cơ quan nhà nước đã sụp đổ từ bên trong, như kim loại bị gỉ. Đó là hình ảnh một người gặm nhấm thể chế của cả một quốc gia từ bên trong. Những câu như vậy, khi ra tòa, đòi hỏi phải được chứng minh. Ai, khi nào, đã làm mục ruỗng thể chế nào, bằng cách nào, tất cả phải có chứng cứ chống đỡ. Bản cáo trạng đưa ra hơn

hai mươi hành vi cụ thể, đánh dấu bằng chữ cái từ "a" đến "s", và gọi đó là chứng cứ. Có mục ghi rõ ngày tháng và số tiền. Có mục lại rộng như "khoảng từ năm 2006 đến năm 2008". Sự rõ ràng và mơ hồ nằm chung trong một văn kiện.

Các cáo buộc gồm ba nhóm: âm mưu khủng bố ma túy, âm mưu nhập khẩu cocaine, và sở hữu súng máy cùng thiết bị phá hoại. Trong các tội danh được ghi trong cáo trạng, riêng âm mưu nhập khẩu cocaine, nếu khối lượng từ năm kilogram trở lên, sẽ chịu mức án tối thiểu 20 năm theo luật Hoa Kỳ.

20 năm.

Trên giấy, các cáo buộc nặng nề và cụ thể. Trọng lượng, loại vũ khí, ngày tháng và số tiền đều được ghi ra. Nhưng ngay giữa lý lẽ ấy có một cụm từ được đóng vào: "Cartel Mặt Trời".

Bản cáo trạng giải thích cụm từ này như sau. Nó viết rằng các nhóm tinh hoa Venezuela tham nhũng vận hành một hệ thống bảo trợ để làm giàu bằng buôn ma túy, và những người đứng trên đỉnh hệ thống ấy được gọi là "Cartel de los Soles, tức Cartel Mặt Trời". Rồi cáo trạng mở ngoặc và thêm một dòng: tên gọi ấy "chỉ biểu tượng mặt trời gắn trên quân phục của các sĩ quan quân đội cấp cao Venezuela".

Biểu tượng mặt trời trên quân phục. Chính đoạn này là mắt xích yếu trong lập luận.

Trong quân đội Venezuela, cấp hiệu của tướng lĩnh có biểu tượng mặt trời. Không phải ngôi sao, mà là mặt trời. Từ thập niên 1990, ở Venezuela, cụm "Cartel Mặt Trời" được dùng như một cách nói quen thuộc để chỉ chung các tướng lĩnh tham nhũng dính vào ma túy. Nó gần với một biệt danh gom những tướng lĩnh tham nhũng cùng loại lại với nhau, chứ không phải tên của một tổ chức tội phạm thống nhất có thủ lĩnh cụ thể, có sơ đồ tổ chức và có trụ sở. Nếu một người là tướng lĩnh tham nhũng, người ta có thể nói ông ta thuộc "Cartel Mặt Trời". Nhưng cụm từ ấy không chỉ một danh sách thành viên của một tổ chức.

Khác biệt này có thể trông nhỏ. Nhưng trong thế giới pháp luật, nó không nhỏ. Muốn truy tố một người là thủ lĩnh của một tổ chức khủng bố ma túy, tổ chức ấy phải tồn tại thật. Tổ chức phải có cấu trúc, và vị trí của người đó trong cấu trúc ấy phải được xác định. Một biệt danh không có cấu trúc như vậy. Một nhóm lỏng lẻo gọi là "các tướng lĩnh tham nhũng" không có thủ lĩnh. Vậy mà sang năm 2025, chính phủ Hoa Kỳ đã tiến hành thủ tục chỉ định "Cartel de los Soles" là tổ chức khủng bố nước ngoài, rồi nêu Maduro là người đứng đầu. Một cách nói quen thuộc trở thành tổ chức; tổ chức ấy có thủ lĩnh; chiến dịch bắt thủ lĩnh ấy được biện minh. Ở điểm khởi đầu của dòng chảy đó là một tờ giấy.

Nhìn vào những nhân vật được bản cáo trạng phác họa, có thể thấy vụ án này không phải là vấn đề của một người. Diosdado Cabello Rondón, thân tín của Maduro; cựu Bộ trưởng Nội vụ Ramón Rodríguez Chacín; vợ ông, Cilia Flores; con trai ông, Maduro Guerra; và cả Héctor Guerrero Flores, thủ lĩnh của tổ chức tội phạm Tren de Aragua (TdA) xuất phát từ các nhà tù

Venezuela. Bản cáo trạng viết rằng những người này đã bắt tay với các tổ chức ma túy vũ trang như Lực lượng Vũ trang Cách mạng Colombia, Quân đội Giải phóng Quốc gia (ELN), Sinaloa Cartel và Zetas để đưa hàng nghìn tấn cocaine vào Hoa Kỳ trong suốt nhiều thập niên. Trong đó cũng có ước tính của Bộ Ngoại giao Hoa Kỳ rằng đến năm 2020, lượng cocaine đi qua Venezuela được cho là từ 200 tấn đến 250 tấn mỗi năm.

Vậy 'cartel' mà bản cáo trạng mô tả là một tổ chức có thật, hay chỉ là một cách nói quen thuộc được khoác lên lớp vỏ của một tổ chức?

Theo bài viết của Common Dreams, một hãng truyền thông có khuynh hướng cấp tiến, chính Bộ Tư pháp Hoa Kỳ trên thực tế đã thừa nhận trước tòa rằng Cartel de los Soles không phải là một tổ chức đơn nhất có thật. Điều từng được gọi trên giấy như một cartel duy nhất, khi đứng trước sự kiểm chứng của tòa án, lại không hiện ra như một thực thể rõ ràng đến vậy. Một khoảng hở đã mở ra giữa cái tên dùng để định danh kẻ thù và thực thể đứng sau cái tên ấy.

Ở điểm này, cách nhìn khác nhau tùy theo từng cơ quan truyền thông. Tài liệu từ phía chính phủ Hoa Kỳ và các bài viết đi theo hướng đó coi Cartel de los Soles là một tổ chức có thật do Maduro cầm đầu. Các bài viết ở phía ngược lại và một số chuyên gia về Venezuela lại cho rằng đó không hẳn là một tổ chức, mà là tên gọi cho một kiểu tham nhũng. Một bên gọi đó là tổ chức, bên kia gọi đó là hiện tượng. Ghi nhận chính sự va chạm giữa hai cách nhìn ấy mới là cách chính xác. Nếu ép nó khép lại theo một phía, sự thật sẽ bị gọt bớt. Có một điều rõ ràng. Ngay giữa lý do được dùng để tiến hành chiến dịch đưa một người sang lục địa khác để trừng phạt, lại đặt một cái tên mà chính thực thể của nó vẫn còn gây tranh cãi.

Vì sao khoảng hở này quan trọng? Muốn đưa một người từ nước khác về để trừng phạt, cần phải rõ người đó là thành viên của cái gì. Nếu ông ta là thủ lĩnh của một tổ chức tội phạm đơn nhất, chiến dịch nhằm tháo dỡ tổ chức đó có thể được biện minh. Nhưng nếu ông ta chỉ là một người bị gom dưới một biệt danh dùng chung để chỉ các tướng lĩnh tham nhũng, câu chuyện sẽ khác. Biệt danh không có thủ lĩnh. Không thể tháo dỡ một cách nói quen thuộc. Thế nhưng chiến dịch lại chỉ đích danh một người ở vị trí chóp bu và đưa người đó ra khỏi nơi ông ta đang ở.

Đến đây, phải đặt cả hai phía lên bàn mới công bằng.

Những sự kiện riêng lẻ ghi trong bản cáo trạng, trọng lượng cocaine bị tịch thu, loại máy bay, số tiền hối lộ, là các cáo buộc cụ thể mà công tố Hoa Kỳ nói rằng họ có chứng cứ. Một phần trong số đó đã được tòa án xác định là có tội. Bản cáo trạng tự nêu rằng hai người họ hàng của Maduro đã bị bồi thẩm đoàn ở New York tuyên có tội vào năm 2016 về tội thông đồng nhập khẩu cocaine. Cũng có đoạn nói rằng cựu lãnh đạo cơ quan tình báo quân đội Venezuela đã nhận tội tại cùng tòa án vào tháng 6 năm 2025 về các cáo buộc khủng bố ma túy và vũ khí. Điều đó không có nghĩa toàn bộ các cáo buộc trên giấy đều là hư cấu.

Phía Maduro lâu nay phản bác rằng bản cáo trạng này là một mục tiêu chính trị nhằm thay đổi chế độ. Lập luận của họ là Hoa Kỳ lấy ma túy làm cái cớ để can thiệp vào dầu mỏ và chính quyền Venezuela. Và nhận xét rằng tên gọi 'Cartel de los Soles' không khớp với thực thể của một tổ chức đơn nhất đã chạm vào phần mềm yếu, chứ không phải phần vững chắc, của lý do ấy. Sức nặng của từng cáo buộc riêng lẻ và sự lỏng lẻo của cái khung gom các cáo buộc ấy dưới một tên tổ chức là hai vấn đề khác nhau. Không được trộn lẫn hai việc đó.

Một chiến dịch đưa một người sang lục địa khác giữa đêm cần một cái tên để biện minh cho nó. Ở Gaza, cái tên ấy là 'đặc vụ Hamas'. Ở Iran, cái tên ấy là 'mối đe dọa hạt nhân'. Ở Venezuela, cái tên ấy là 'Cartel de los Soles'.

Khi tên gọi của lý do được định sẵn trước cả thực thể thật, rồi chiến dịch bắt đầu lăn bánh dưới cái tên ấy, thứ biến mất một lần nữa là nơi đặt trách nhiệm. Ai đã xác định người này là kẻ thù? Việc xác định ấy đã được kiểm chứng đến đâu? Phần chưa được kiểm chứng sẽ do ai gánh?

Giữa tờ giấy của năm 2020 và chiến dịch của năm 2026 có khoảng cách sáu năm. Khoảng cách này quan trọng. Năm 2020, khi bản cáo trạng được nộp, chưa có cách nào bắt được Maduro. Những cáo buộc trên giấy vẫn nằm trên giấy. Muốn bắt ông ta, phải biết ông ta ở đâu, phải đến được nơi đó, và phải xuyên qua lớp phòng thủ ở đó. Năm 2020, cả ba việc ấy đều khó. Nhưng đến năm 2026, cả ba đều đã làm được. Năm tháng truy vết cho biết ông ta ở đâu, máy bay từ hai mươi căn cứ đã đến được nơi đó, và cuộc tấn công mạng đã xuyên qua hệ thống phòng thủ. Lý do đã được chuẩn bị từ sáu năm trước, nhưng năng lực thi hành lý do ấy chỉ mới có gần đây. Câu nói của Karp, rằng đó là việc 'hai năm trước, ba năm trước chưa thể làm', vang lại ở đây. Tờ giấy đã được chuẩn bị từ lâu, và thứ biến tờ giấy ấy thành hiện thực là công nghệ mới vừa được trang bị.

Nếu nghĩ ngược lại trình tự này, một bức tranh bất an hiện ra. Có phải họ bắt ông ta vì giờ đã có thể bắt được? Có phải chiến dịch được tung ra không phải vì lý do đã đủ chín, mà vì công cụ thi hành đã đủ chín? Tôi sẽ không khẳng định. Nhưng khoảng cách giữa lý do và năng lực cho thấy một thời kỳ đã đến, khi năng lực kéo lý do lại gần mình. Khi những việc có thể làm tăng lên, tiếng nói hỏi rằng có được phép làm hay không sẽ nhỏ đi.

Tên gọi là do con người đặt ra. Trong bản cáo trạng năm 2020, bàn tay con người đã viết cụm từ 'Cartel de los Soles'. Nhưng dưới cái tên ấy, việc thu hẹp mục tiêu, nối các tuyến di chuyển, và sắp xếp thứ tự xuất kích giờ đây không còn chỉ có con người.

## 3. Vết nứt: Anthropic và Palantir, cùng mâu thuẫn

Hãy hình dung một màn hình.

Đó là một màn hình trong phòng tác chiến. Trên đó, thông tin rải rác được gom lại một nơi. Những gì vệ tinh nhìn thấy, những gì tín hiệu bắt được, những gì đội xâm nhập đã ghi gửi về trong năm tháng, tất cả chồng lên nhau trên một tấm kính. Palantir gọi điều này là 'một tấm

kính duy nhất (plane of glass)'. Nguồn thông tin thì mỗi thứ một khác. Có thứ quá nhạy cảm nên không ai được xem, có thứ mọi người đều phải xem. Cũng có thông tin mà các nước đồng minh cùng nhìn vào. Đưa tất cả những thứ ấy lên một màn hình trong khi vẫn giữ được bảo mật, đó là thứ Palantir bán.

Trong lúc chiến dịch Maduro vận hành, Claude của Anthropic hiện lên trên tấm kính ấy. Đây là thông tin từ Semafor.

Có lẽ nên bắt đầu bằng việc Palantir là công ty bán thứ gì. Trong một cuộc phỏng vấn truyền hình, Alex Karp, đồng sáng lập kiêm CEO của công ty, đã mô tả nền tảng của công ty mình như sau.

"Nền tảng Maven của Palantir lấy các mô hình ngôn ngữ lớn có giá trị và biến chúng thành thứ thật sự có sức sát thương và hữu dụng trên chiến trường."

Biến mô hình ngôn ngữ lớn thành thứ có sức sát thương trên chiến trường. Karp không nói vòng vo. Trong cùng cuộc phỏng vấn, ông nói mục tiêu của công ty là đưa binh sĩ Mỹ trở về nhà an toàn, và để làm được vậy thì phải "tháo rã" kẻ địch. Ông trực tiếp nhắc đến Venezuela và nói rằng Mỹ đã cho thấy "năng lực tháo rã những thứ như thế này".

"Nếu nhìn vào thành quả của Mỹ, nhờ lòng can đảm và sự sáng tạo của người Mỹ, và tôi muốn tin là nhờ cả công nghệ của chúng tôi, tổn thất phía chúng ta thấp, chính xác và được kiểm soát. Như quý vị đã thấy ở Venezuela, đó là năng lực hoàn chỉnh để tháo rã những thứ như thế này. Đây là việc 2 năm trước, 3 năm trước chưa thể làm được."

Việc 2 năm trước chưa thể làm được. Đó là lời do chính Karp nói ra. Nghĩa là trong vài năm gần đây, việc thu hẹp mục tiêu, nối các tuyến di chuyển và vận hành chiến dịch trên một màn hình đã trở nên khả thi. Thứ hiện lên trên màn hình ấy chính là Claude của Anthropic.

Karp giải thích Maven làm gì bằng ví dụ Ukraine. Ông nói khó khăn thật sự không chỉ là đưa drone đến điểm mục tiêu, mà là xuyên qua các hệ thống radar tiên tiến nằm giữa đường. Nối những hệ thống cũ rời rạc với hệ thống mới, gom thông tin có cấp độ bảo mật khác nhau lên một màn hình, rồi đặt trí tuệ nhân tạo và học máy lên trên đó để lọc ra thứ cần nhìn. Ông giải thích rằng đó chính là Maven. Còn với Iran, ông nói "phức tạp hơn 100 lần". Lý do là phòng không của đối phương dày đặc đến mức ấy.

Karp cũng phân hạng trí tuệ nhân tạo. Ông gọi loại trí tuệ nhân tạo vô dụng trên chiến trường là "mã AI lỏng lẻo". Ông nói những thứ như vậy không được dùng ở bất kỳ chiến trường thực tế nào. Theo cách phân biệt của ông, mô hình ngôn ngữ lớn tự thân là công cụ dựa vào xác suất. Nó có thể dùng được trong những việc chỉ cần đúng 51 phần trăm, như phán đoán đầu tư, nhưng mức đó không đủ cho việc đặt vũ khí chính xác lên đầu kẻ địch và đưa binh sĩ Mỹ trở về an toàn. Vì vậy, lập luận của ông là chỉ mô hình ngôn ngữ lớn thôi thì chưa đủ; nó chỉ có ích khi có nền tảng và sự quản lý để vận hành đúng cách.

Nếu đảo ngược câu nói ấy, vấn đề sẽ thành thế này. Những mô hình ngôn ngữ lớn như Claude của Anthropic tự thân không phải là vũ khí. Việc biến chúng thành vũ khí là công việc của Palantir: nối chúng với thông tin chiến trường, đặt lớp bảo mật lên chúng, rồi đưa chúng lên màn hình tác chiến. Chính Karp đã gọi công việc đó là "làm cho mô hình ngôn ngữ lớn có năng lực sát thương". Vậy năng lực sát thương đến từ mô hình, hay đến từ bàn tay đưa mô hình ấy lên chiến trường? Công ty tạo ra mô hình có thể nói rằng mô hình của mình không phải vũ khí. Công ty đưa mô hình lên chiến trường có thể nói rằng mình chỉ lắp đặt công cụ, còn cò súng là do con người bóp. Năng lực sát thương sinh ra trong khoảng trống giữa hai công ty, nhưng không công ty nào gọi khoảng trống ấy là của mình.

Vết nứt bắt đầu trước màn hình ấy.

Theo bản tin của Semafor, trong một buổi kiểm tra định kỳ, phía Anthropic đã bày tỏ sự khó chịu trước việc dùng mô hình của họ cho mục đích quân sự. Sau đó, một lãnh đạo của Palantir đã báo cáo sự khó chịu ấy lên Lầu Năm Góc. Một vết rạn xuất hiện giữa công ty làm ra vũ khí và công ty làm ra trí tuệ nhân tạo đặt trên vũ khí đó. Một bên nói sẽ dùng nó trên chiến trường, bên kia vạch ranh giới với cách dùng ấy.

Cần nhìn cảnh này thật chậm. Anthropic là công ty tạo ra Claude. Palantir là công ty đặt Claude lên nền tảng của mình rồi bán cho quân đội. Hai công ty từng là đối tác. Nhưng trong buổi kiểm tra định kỳ, tức là nơi hai bên cùng xem công việc có vận hành tốt hay không, phía tạo ra mô hình đã tỏ ra bất an trước cách nó được dùng. Người làm ra thứ đó đã khựng lại khi nhìn thấy nó được đem đi dùng ra sao.

Phản ứng của Bộ trưởng Quốc phòng Pete Hegseth được thuật lại là rất rõ. Ý của ông là sẽ không dùng một trí tuệ nhân tạo ngăn cản việc tiến hành chiến tranh. Theo các bản tin, điều này đã trở thành ngòi nổ cho việc chỉ định rủi ro chuỗi cung ứng và đưa vào danh sách đen sau đó. Đó là tín hiệu rằng công cụ nào do dự trong việc hỗ trợ chiến tranh thì sẽ bị gạt sang một bên. Chính sự do dự đã trở thành lý do loại bỏ.

Ở đây có một mâu thuẫn. Và đó là mâu thuẫn được viết trên giấy.

Chính sách sử dụng của Anthropic ghi rõ rằng công ty cấm dùng mô hình Claude của mình cho vũ khí và giám sát. Nguyên tắc công ty đưa ra là có thể tạo ra nó, nhưng không được dùng nó như vũ khí. Thế nhưng mô hình của công ty ấy lại xuất hiện trên tấm kính trong phòng tác chiến nơi chiến dịch Maduro đang vận hành. Chính sách cấm vũ khí và giám sát, nhưng mô hình lại ở ngay giữa một chiến dịch quân sự.

Nếu công ty tự khâu lại mâu thuẫn này, nó vẫn có thể được khâu lại. Mô hình của chúng tôi không trực tiếp bóp cò; nó không tham gia vào bước cuối cùng của việc sát thương; chính sách của chúng tôi vẫn được tuân thủ. Cũng có thể có lập luận rằng phân tích quân sự khác với hành vi sát thương. Sắp xếp thông tin và nhắm vào con người không phải là một.

Nhưng câu hỏi mà cuốn sách này luôn bám theo không nằm ở bước cuối cùng. Nó nằm ở toàn bộ những bước phía trước: thu hẹp mục tiêu, nối các mẩu thông tin, chuẩn bị màn hình để con người đóng dấu phê chuẩn. Chúng ta đã nghe lời biện minh rằng Lavender ở Gaza không giết người, nó chỉ đưa người ta vào danh sách. Giữa việc đưa vào danh sách và việc giết người có 20 giây của sĩ quan tình báo, và 20 giây ấy chỉ là thời gian để đóng dấu. Nói rằng mình không bóp cò không hề mâu thuẫn với sự thật rằng mình đã tham gia vào mọi bước trước khi ngón tay đặt trước cò súng.

Karp đã tự giải tỏa căng thẳng ấy theo cách của mình trong một cuộc phỏng vấn khác. Ông nói về Dario Amodei của Anthropic như một người bạn, một đối thủ, đồng thời là người có quan điểm khác với mình. Ông cũng nói rằng ngoài đời riêng họ vẫn thân thiết. Rồi ông châm chọc rằng các công ty mô hình ngôn ngữ lớn thường tin rằng một ngày nào đó mọi vấn đề sẽ tự nhiên được giải quyết. Ý ông là bản thân ông thì khác.

"Tôi tin rằng chúng ta cần thiên đường ngay trên mặt đất này, chứ không phải thiên đường sau 20 năm nữa."

Và những gì ông nói về trí tuệ nhân tạo đặt ranh giới với việc dùng vũ khí còn trực diện hơn. Lập luận của ông là kẻ thù sẽ không đặt ra những ranh giới như vậy.

"Nếu không dùng nó trên chiến trường, chắc chắn Iran sẽ dùng. Các vị nghĩ họ không thể mua những sản phẩm như thế trên mạng sao."

Ở đây, vị trí của Karp rất rõ. Ông đã chọn một phía. Phía Mỹ, phía phương Tây, ông lặp đi lặp lại điều đó trong suốt cuộc phỏng vấn. Ông nói công ty của mình gây tranh cãi vì "cung cấp sức mạnh cho mọi cường quốc phương Tây đang trong chiến tranh". Trong lập luận của ông, nếu đó là công cụ mà kẻ thù sẽ dùng, thì chúng ta cũng phải dùng. Ông không dao động ở điểm này. Nhắc đến các nước thù địch với Mỹ, ông còn nói rằng "nhiều kẻ muốn làm hại nước Mỹ trên chiến trường rốt cuộc sẽ chết". Theo ông, đó là nhờ năng lực thu thập thông tin và nắm tình hình chiến trường trước đối phương.

Lập luận của Karp nhất quán ở một điểm. Ông cho rằng bên buông công cụ khỏi tay sẽ thua. Nếu kẻ thù không do dự, chúng ta cũng không thể do dự. Trong lập luận ấy, chính sách cấm dùng vũ khí không còn là đạo đức, mà trở thành điểm yếu. Nếu người vạch ranh giới thất bại vì chính ranh giới ấy, thì việc vạch ranh giới trở thành một thứ xa xỉ.

Karp phân biệt rõ bên trong và bên ngoài biên giới. Trong một podcast, ông nói công dân Mỹ có quyền tự do biểu đạt và quyền riêng tư. Vì vậy, ông cũng thấy khó chấp nhận việc tùy tiện dùng những công cụ như vậy trong hoạt động thực thi pháp luật bên trong nước Mỹ. Nhưng ông khẳng định rằng kẻ thù ở Iran muốn giết người Mỹ thì không có những quyền ấy.

"Những kẻ thù Iran muốn giết chúng ta không có quyền đó. Tôi chưa bao giờ tin vào việc mở rộng các quyền của chúng ta cho một nước ngoài thù địch."

Trong cách phân chia ấy, Venezuela nằm ở phía nào. Maduro là một nhà cai trị nước ngoài. Theo khung nhìn của Karp, ông là kẻ thù nằm ngoài phạm vi quyền. Nhưng theo khung nhìn của phía Maduro, ông là nguyên thủ của một quốc gia có chủ quyền. Cùng một con người, một bên gọi là kẻ thù không có quyền, bên kia gọi là nguyên thủ quốc gia bị bắt cóc. Đường ranh mà Karp vạch ra rất gọn. Bên trong và bên ngoài, chúng ta và kẻ thù. Vấn đề là ai vạch đường ấy, và vạch ở đâu. Khoảnh khắc Maduro bị đặt về phía kẻ thù, mọi việc nhằm bắt ông đều trở nên chính đáng. Chính hành vi vạch đường đã là một nửa chiến dịch.

Karp cũng nói rằng những công cụ này khiến việc nói dối trở nên khó hơn. Nếu không tạo ra được giá trị nào, hoặc nếu có tham nhũng, thì sẽ không thể che giấu được nữa. Đây là một lập luận đáng nhìn kỹ. Logic ở đây là công nghệ gom thông tin vào một màn hình sẽ làm sự thật lộ ra. Nhưng cùng công nghệ ấy đã không ngăn được việc các căn cứ dùng để quy một người thành kẻ thù sụp đổ trước tòa. Công nghệ gom thông tin lại, nhưng con người quyết định sẽ đặt tên gì lên trên thông tin ấy. Cái tên 'Cartel de los Soles' không xuất hiện từ màn hình, mà từ giấy tờ; không xuất hiện từ thuật toán, mà từ bàn tay của công tố.

Trong những cuộc phỏng vấn ấy, Karp cũng để lộ một nỗi sợ lớn hơn. Ông nhìn thời đại này như một thế giới đang đi tới chỗ chỉ còn Mỹ và Trung Quốc. Đó là một thế giới chia thành những người có trí tuệ nhân tạo và những người không có, một thế giới nơi việc ai trong hai bên sẽ định đoạt trật tự toàn cầu được phân xử. Ông nói mình không muốn làm hại Trung Quốc, chỉ muốn Mỹ thắng. Trong khung cạnh tranh khổng lồ ấy, chiến dịch bắt một người tên Maduro chỉ là một nước đi nhỏ. Nhìn bằng mắt Karp, Venezuela là một ô trên bàn cờ lớn hơn. Nhưng trong ô ấy có việc bắt giữ một con người, có một lý do mờ nhạt về thực chất, và có một màn hình nơi một mô hình mang chính sách không được dùng cho vũ khí đang hiện lên. Logic của bàn cờ lớn che lấp mâu thuẫn trong ô nhỏ. Nếu có thể thắng, thì đừng xét nét những mâu thuẫn nhỏ. Cuốn sách này nhìn vào ô nhỏ ấy. Không nhìn thắng thua của bàn cờ lớn, mà nhìn trách nhiệm đã biến mất trong một ô của bàn cờ đó.

Rồi Karp kéo vào một bức tranh lớn hơn nữa. Trong một cuộc phỏng vấn, ông từng ví thời đại xoay quanh vũ khí trí tuệ nhân tạo này với 'khoảnh khắc Oppenheimer' (Oppenheimer moment). Giống như nhà vật lý đã chế tạo bom hạt nhân, các kỹ sư công nghệ ngày nay cũng đang cầm trong tay những công cụ mạnh mẽ. Sau khi tạo ra hạt nhân, Oppenheimer bị sức nặng của nó đè lên. Kết luận của Karp không phải là dừng lại trước sức nặng ấy, mà là tiến lên phía trước. Theo đuổi, tinh luyện, quản lý, nhưng không buông khỏi tay.

Từ đây, hai công ty rẽ sang hai hướng.

Một bên muốn tạo ra công cụ nhưng vạch ranh giới cho cách dùng nó, bên kia cho rằng nếu vạch ranh giới ấy thì sẽ thua kẻ thù. Cả hai đều là công ty làm trí tuệ nhân tạo. Cả hai đều biết công nghệ ấy nguy hiểm. Một bên vì nguy hiểm đó mà cấm dùng cho vũ khí trong chính sách của mình, bên kia lại lập luận rằng chính vì nguy hiểm đó nên phải dùng nó như vũ khí. Họ nhìn

cùng một mối nguy và đi tới hai kết luận trái ngược.

Không có cách nào vá kín mâu thuẫn này một cách sạch sẽ. Nếu chỉ nói Anthropic đúng, ta không thể giải thích vì sao mô hình của họ lại hiện trên màn hình của một chiến dịch quân sự. Nếu chỉ nói Palantir đúng, lời của một công ty ghi trong chính sách rằng không được dùng cho vũ khí sẽ chẳng còn trọng lượng nào. Tôi không định vá lại mọi chuyện bằng câu nói rằng cả hai bên đều có lý. Tôi muốn phơi bày nguyên trạng mâu thuẫn của cả hai. Chính sách cấm vũ khí, còn trong thực tế, một mô hình mang chính sách ấy đã ở cạnh vũ khí. Hai câu này cùng lúc đều đúng.

Vào đêm Maduro bị bắt, trên tấm kính trong phòng tác chiến, mâu thuẫn của hai công ty ấy hiện lên cùng lúc. Một bên là chính sách ghi rằng không được dùng cho vũ khí; bên kia là thực tế một mô hình mang chính sách ấy lại đang nằm giữa một chiến dịch quân sự. Cả hai chồng lên nhau trên cùng một màn hình.

Vậy tôi xin hỏi. Trong lúc lý do coi Maduro là kẻ thù sụp đổ trước tòa, các công ty đã làm ra công cụ thu hẹp mục tiêu và nối các tuyến di chuyển dưới danh nghĩa ấy đang đứng ở đâu. Người làm ra công cụ có chịu trách nhiệm về việc công cụ ấy nhắm vào điều gì không, hay chỉ cần một dòng trong văn bản chính sách là thoát khỏi trách nhiệm ấy.

Việc hỏi trách nhiệm ngày càng khó hơn. Bên công tố, nơi đặt ra danh nghĩa, nói rằng trên giấy có cáo buộc. Quân đội thực hiện chiến dịch nói rằng họ làm theo mệnh lệnh. Palantir, bên dựng màn hình, có thể nói rằng họ chỉ bán công cụ. Anthropic, bên làm ra mô hình, có thể nói rằng chính sách đã cấm dùng cho vũ khí. Ai cũng có thể nói mình đã làm phần việc của mình ở vị trí của mình. Nhưng với toàn bộ sự việc theo dõi một người suốt năm tháng rồi lôi ông ta đi giữa đêm, không ai nhận trọn đó là việc của mình.

Đây là điều cuốn sách này gọi là chỗ trống. Trách nhiệm không biến mất. Trách nhiệm bị xé nhỏ, mỗi bên chỉ cầm phần của mình. Bên công tố cầm phần cáo buộc, quân đội cầm phần mệnh lệnh, Palantir cầm phần công cụ, Anthropic cầm phần chính sách. Từng phần riêng lẻ trông có vẻ không có gì để bắt bẻ. Nhưng nơi ghép các phần ấy lại, nơi phải chịu trách nhiệm cho toàn bộ việc một con người bị đưa sang lục địa khác, thì trống không. Chuỗi càng dài, càng ít người gọi việc xảy ra ở cuối chuỗi ấy là việc của mình.

Khi một sĩ quan tình báo ở Gaza đóng dấu trong 20 giây, anh ta nghĩ rằng mình chỉ làm theo điều màn hình bảo. Người làm ra màn hình nghĩ rằng mình chỉ sắp xếp thông tin. Trước tấm kính ở Venezuela, chuyện tương tự cũng diễn ra. Màn hình gom thông tin, công tố gắn tên, quân đội làm theo lệnh. Người ở mỗi bước chỉ nhìn bước của mình. Con mắt nhìn toàn cảnh chỉ nằm trong màn hình, còn màn hình thì không chịu trách nhiệm.

Cần ghi thêm một điều nữa thì mới cân bằng. Việc Anthropic bộc lộ sự khó chịu trước việc dùng vào quân sự, tự thân nó cũng có nghĩa là ít nhất đã có một công ty biết khựng lại. Ở

những chương khác mà cuốn sách này đã lần theo, ngay cả sự khựng lại ấy cũng khó thấy. Bàn tay làm ra hệ thống ở Gaza, bàn tay cho drone ở Ukraine cất cánh, đều không công khai ngập ngừng trước cách người ta dùng thứ mình tạo ra. Việc công ty làm ra mô hình nói ra sự bất an trong một buổi kiểm tra định kỳ là một vết nứt nhỏ, nhưng vẫn là vết nứt. Vấn đề nằm ở chỗ vết nứt ấy không được tiếp nhận như một tín hiệu đạo đức, mà bị xử lý như một lý do loại bỏ. Công ty do dự không được khen, mà trở thành ngòi nổ cho danh sách đen. Trong một thế giới nơi sự ngập ngừng trở thành điểm yếu, công ty tiếp theo sẽ học trước cách không ngập ngừng.

Đèn ở Caracas đã sáng trở lại. Maduro đã đứng trước tòa án ở New York. Nhưng trước màn hình đã đưa ông ta tới tòa án ấy, trước tấm kính nối thông tin và thu hẹp mục tiêu ấy, vẫn chưa có ai đứng ra.

Chỗ trống ấy vẫn còn nguyên ở cuối chương này.

## 4. Là mô hình, hay là bàn tay: công nghệ biến mô hình ngôn ngữ lớn thành vũ khí

Tôi đã viết ở phần trước rằng sức sát thương sinh ra từ khoảng trống giữa hai công ty. Vậy trong khoảng trống ấy, điều gì đang xảy ra?

Hãy bắt đầu bằng một cảnh trong phòng tác chiến. Có một màn hình. Trên đó, tên người, tọa độ và thời gian hiện lên thành từng dòng. Có dòng được khoanh bằng ô màu vàng, có dòng gắn dấu hỏi. Màn hình ấy do một công ty tạo ra, còn việc đọc các dòng và ô trên màn hình rồi đoán dòng tiếp theo là việc của mô hình thuộc một công ty khác. Một người ngồi trước màn hình và đóng dấu phê chuẩn. Trong bức tranh này, sức sát thương nằm ở đâu? Ở màn hình, ở mô hình, hay ở bàn tay cầm con dấu?

Ở cuối phần trước, chúng ta đã thấy mâu thuẫn giữa hai công ty. Mô hình của công ty ghi trong chính sách rằng không được dùng cho vũ khí lại xuất hiện trên tấm kính trong phòng tác chiến, nơi chiến dịch Maduro đang vận hành. Mâu thuẫn ấy là khoảng hở giữa văn bản chính sách và thực tế. Phần này lần theo khoảng hở ấy bằng tay. Một mô hình bị cấm dùng cho vũ khí trong chính sách rốt cuộc đã đi qua những bàn tay nào để xuất hiện trên màn hình phòng tác chiến? Những bàn tay ấy đã thêm gì vào mô hình, gọt bỏ gì khỏi mô hình, để một bộ não viết văn biến thành công cụ chỉ vào mục tiêu?

Phần này bước vào bên trong khoảng trống ấy. Giữa bàn tay tạo ra mô hình và bàn tay đưa mô hình lên chiến trường, chúng ta sẽ nhìn kỹ xem chuyện gì đã xảy ra để một dòng chữ trở thành mục tiêu. Sẽ có kỹ thuật khó. Nhưng nếu đi chậm, nó không khó. Điểm chính có thể rút lại trong một câu. Mô hình là công cụ viết chữ, còn thứ biến công cụ ấy thành thứ dùng được trên chiến trường là các thiết bị được dựng quanh nó. Chúng ta sẽ giữ lấy câu này để đi tiếp. Mô hình là một chuyện, thiết bị là một chuyện khác. Hãy xem điều gì sinh ra ở nơi hai thứ ấy gặp nhau.

Mô hình không phải là vũ khí, tích hợp mới biến nó thành vũ khí

Giám đốc công nghệ của Palantir, Shyam Sankar, đã nói gọn trong một câu về thứ công ty của ông bán. Theo sự kiện của Hudson Institute và bài đưa tin của The Register, ông nói như sau.

"Phần mềm là hệ thống vũ khí quan trọng nhất. Đó là cách chúng tôi chuyển giao sức sát thương áp đảo."

Cách nói rằng phần mềm là một hệ thống vũ khí nghe khá lạ. Khi nói đến vũ khí, người ta thường nghĩ đến súng hoặc tên lửa. Thứ cầm trong tay để ngắm bắn, thứ bay đi rồi phát nổ. Nhưng Sankar không gọi khối sắt thép nhìn thấy được là vũ khí. Ông gọi chương trình gom các khối sắt thép ấy lại và khiến chúng chuyển động cùng nhau là vũ khí.

Hình ảnh ông đưa ra là "từ sàn nhà máy đến chiến hào". Ở một đầu là nhà máy. Đó là nơi chế tạo máy bay chiến đấu và sản xuất đạn dược. Ở đầu kia là chiến hào. Đó là nơi người lính đối mặt với kẻ địch. Ở giữa, máy bay chiến đấu, máy bay ném bom, drone và đạn dược được đặt thành hàng. Nếu những thứ rời rạc này vận hành riêng lẻ, chúng không có sức mạnh nào đáng kể. Máy bay ném bom không biết phải đánh vào đâu, drone không biết phải nhìn cái gì, đạn dược không biết sẽ được dùng cho mục tiêu nào. Nếu ai bắn, bắn ở đâu, bắn cái gì bị lệch nhau, thiết bị đắt tiền đến mấy cũng chỉ là sắt vụn. Ghép chúng lại trên một màn hình và khiến chúng chuyển động cùng một nhịp. Khớp trên một màn hình nơi máy bay ném bom sẽ đánh, thứ drone sẽ nhìn, và nơi đạn dược sẽ đi tới. Theo Sankar, sát thương sinh ra từ đó. Vũ khí không phải là khối sắt thép, mà là sợi dây nối các khối sắt thép ấy.

Ở đây, lời Karp nói lại vang lên. Ông nói Maven "lấy các mô hình ngôn ngữ lớn có giá trị rồi biến chúng thành thứ có sát thương và hữu dụng trên chiến trường trong thực tế". Khung câu giống nhau. Mô hình tự thân không có sát thương, và phải có thứ gì đó biến đổi nó cho phù hợp với chiến trường.

Trong câu nói ấy có một lời thú nhận ẩn bên trong. Đó là cụm "lấy rồi". Karp không nói rằng ông tự tạo ra mô hình ngôn ngữ lớn. Ông nói rằng họ lấy nó từ đâu đó. Nơi lấy ấy là các công ty như Anthropic. Có công ty riêng tạo ra mô hình, rồi lại có công ty khác lấy mô hình ấy và sửa nó cho phù hợp với chiến trường. Công ty tạo ra mô hình không nói rằng mình tạo ra sát thương. Công ty sửa mô hình không nói rằng mình tạo ra mô hình. Từ sát thương được đặt lên trên một cụm "lấy rồi" nằm giữa hai công ty đó.

Karp cũng giải thích việc sửa đổi ấy là gì. Theo tường thuật của The Register, ông gọi mô hình ngôn ngữ lớn là một công cụ "không thể dự đoán". Cần ngẫm kỹ cụm không thể dự đoán. Mô hình viết văn bản trả lời hơi khác nhau mỗi lần, ngay cả với cùng một câu hỏi. Hôm qua nó trả lời thế này, hôm nay nó trả lời thế khác. Cho nó xem cùng một bức ảnh hai lần, lần đầu nó có thể nói là xe tải, lần sau lại nói là xe van. Đây là bản tính của công cụ viết văn bản. Mô hình không phải là chiếc máy lấy ra một đáp án đã định sẵn, mà là chiếc máy mỗi lần lại dựng nên

một câu trả lời có vẻ hợp lý.

Không thể dùng nguyên xi một công cụ như vậy để chọn mục tiêu cho tên lửa ngắm bắn. Một khi mục tiêu được xác định, con người sẽ chết. Không thể giao việc đó cho một công cụ mỗi lần lại trả lời khác nhau. Theo giải thích của Karp, nền tảng AIP của Palantir bao quanh mô hình ấy bằng tính minh bạch, bảo mật và lan can kiểm soát, để khi mô hình không thể dự đoán hỗ trợ nhận diện mục tiêu, nó trở nên dùng được. Lan can kiểm soát giống như tay vịn dựng hai bên đường. Đó là hàng rào ngăn mô hình đi chệch khỏi đường. Tính minh bạch là cho con người nhìn vào được vì sao mô hình trả lời như vậy, còn bảo mật là khóa lại để câu trả lời ấy không rò rỉ ra ngoài. Bản thân mô hình vẫn giữ nguyên, nhưng quanh nó được dựng lan can và gắn ổ khóa. Lời Karp rất rõ. Mô hình để nguyên thì không dùng được. Phải bao quanh nó rồi mới dùng được.

Có một thông cáo chính thức nói rằng Anthropic và Palantir đã bắt tay nhau. Thông cáo đăng trên Business Wire, dịch vụ phân phối thông cáo báo chí, viết rằng AIP "operationalize" Claude, tức đưa Claude vào vận hành tác chiến. Từ đưa vào vận hành tác chiến là then chốt. Nó có nghĩa là đưa mô hình ra khỏi phòng thí nghiệm, cắt gọt và chỉnh sửa để đặt nó vào hiện trường tác chiến. Mô hình vẫn là mô hình ấy, nhưng nơi mô hình được đặt đã chuyển sang phòng tác chiến. Cùng là Claude, nhưng khi ở cửa sổ trò chuyện và khi ở trên màn hình phòng tác chiến, công việc nó làm khác nhau.

Nếu tách rõ hai công ty này làm gì và không làm gì, toàn bộ mục này sẽ hiện ra rõ hơn. Anthropic nặn ra mô hình. Họ tạo ra bộ não đọc và viết. Palantir xây chiếc ghế, chiếc bàn và căn phòng để bộ não ấy ngồi vào. Họ sắp xếp thứ bộ não phải nhìn rồi đặt lên bàn, và gắn ổ khóa vào căn phòng để câu trả lời của bộ não không rò rỉ ra ngoài. Chỉ có bộ não thì chẳng làm được gì. Chỉ có căn phòng cũng chẳng làm được gì. Hai thứ phải gặp nhau thì công việc mới thành. Khi Karp nói "biến mô hình ngôn ngữ lớn thành thứ có sát thương", điều ông chỉ vào chính là cuộc gặp ấy. Không phải phía tạo ra bộ não, mà là phía xây căn phòng đã tự miệng nói ra chữ sát thương.

Giải phẫu tầng tích hợp, biến thông tin hỗn tạp thành dạng máy có thể đọc

Muốn đưa mô hình ra chiến trường, trước hết phải biến thông tin chiến trường thành dạng mô hình có thể đọc. Đây là trái tim của việc tích hợp.

Thông tin chiến trường rất hỗn tạp. Có video do drone quay, có tín hiệu radar thu được, có tài liệu do con người viết tay. Hình dạng của chúng đều khác nhau. Con người nhìn những thứ ấy bằng mắt rồi nối chúng lại trong đầu. Nhưng máy móc không làm được như vậy. Máy móc cần một dạng đã được sắp xếp.

Palantir đảm nhiệm việc sắp xếp đó. Theo giải thích của trang bình luận kỹ thuật không chính thức Towards AI, Palantir đặt các thông tin rời rạc ấy lên một nền tảng gọi là "ontology". Từ

ontology nghe khó, nhưng nếu vẽ ra thì dễ hiểu. Hãy nghĩ đến một thư viện. Nếu sách bị chất bừa bãi, ta không thể tìm được gì. Khi gắn mã phân loại, gom theo tác giả, chia ô theo chủ đề, ta mới tìm được sách. Ontology là việc gắn những mã phân loại như vậy cho thông tin chiến trường. Đây là người, kia là xe, kia là tòa nhà. Bằng cách tạo một bản đồ gồm các "đối tượng đã được gán loại" như thế, nó trải đường để mô hình có thể suy luận trên đó.

Bài giải thích ấy so sánh cách này với một cách khác. Hiện nay có một phương pháp thường được dùng gọi là RAG. Đó là cách mô hình tìm các văn bản liên quan trước khi trả lời rồi dùng chúng. Nghĩa là tìm kiếm và mang văn bản về. Khi hỏi mô hình điều nó không biết, trước tiên ta tìm các đoạn liên quan trong thư viện, trải chúng ra trước mô hình, rồi để mô hình nhìn vào đó mà trả lời. Towards AI phân biệt cách làm của Palantir với phương pháp này và gọi nó là "Ontology-Augmented Generation", OAG. Nghĩa là thay vì kéo văn bản về, nó kéo về bản đồ các đối tượng đã được phân loại.

Vì sao sự khác biệt này quan trọng? Văn bản mờ. Cùng một câu có thể được đọc khác nhau tùy người đọc. Câu "chiếc xe tải màu đen đi về phía bắc" là đủ với con người, nhưng lại mơ hồ với máy móc. Đó là xe tải nào, phía bắc là tọa độ nào, chiếc xe tải ấy có phải cùng chiếc đã thấy trước đó hay không, những điều đó không hiện rõ trong câu chữ. Trái lại, đối tượng đã phân loại thì rõ. Đối tượng xe tải được gắn một mã riêng, mã ấy nối với chiếc xe tải đã thấy hôm qua, vị trí được đóng vào tọa độ. Mô hình không lang thang trên lớp chữ mờ, mà tìm đường trên bản đồ các đối tượng rõ ràng. Đó là khác biệt giữa việc vào thư viện tìm sách để đọc và việc lật những thẻ chỉ mục đã được sắp xếp sẵn. Trên chiến trường, khác biệt ấy trở thành tốc độ, và tốc độ chính là sức sát thương.

Trong cùng bài giải thích còn có một câu nữa. Đáng nhìn kỹ. Bài viết nói rằng về mặt cấu trúc, mô hình ngôn ngữ lớn đó bị chặn không cho tự thực hiện hành vi tấn công vật lý. Tiếng Anh viết là "architecturally prohibited from kinetic action". Nếu ví với kiến trúc, nó giống như giữa hai căn phòng ngay từ đầu đã không có cửa. Mô hình có thể đọc và sắp xếp thông tin, nhưng không có cánh cửa nào dẫn sang căn phòng có cò súng. Tuy nhiên, lời giải thích này đến từ một bài bình luận kỹ thuật không chính thức. Cần đối chiếu với tài liệu gốc của công ty. Tôi sẽ không biến suy đoán thành sự thật.

Câu này quan trọng vì nó chạm vào lời biện minh của hai công ty đã nói ở mục trước. Mô hình đã bị chặn để không thể tấn công. Vì vậy công ty tạo ra mô hình có thể nói rằng mô hình của mình không giết người. Nhưng thứ bị chặn chỉ là cánh cửa cuối cùng. Tất cả các hành lang đưa mục tiêu đến trước cánh cửa ấy đều có mô hình cùng đi qua.

Một hệ thống cho thấy hành lang này trong thực tế trông như thế nào. Đó là Maven Smart System. Theo phân tích của think tank Mỹ CSIS, khi hệ thống này được triển khai tại Bộ Tư lệnh Trung tâm, nó đã gom 179 nguồn dữ liệu lại một chỗ. Hình ảnh do drone ghi lại, radar khẩu độ tổng hợp, radar mặt đất, tình báo tín hiệu, tài liệu viết tay. 179 dòng dữ liệu thuộc đủ loại khác

nhau cùng chảy vào một màn hình. Trên đó, chức năng nhận dạng mục tiêu tự động tìm ra mục tiêu. Mục tiêu được tìm thấy hiện lên trên màn hình trong khung màu vàng. Công việc trước đây con người phải nhìn kỹ từng đoạn video để tìm từng mục tiêu, nay máy đánh khung và chỉ ra trước.

CSIS truyền đạt hiệu quả ấy bằng con số. Theo lời một quan chức trong một cơ quan, chỉ riêng thị giác máy tính đã làm năng lực nhắm mục tiêu tăng khoảng 10 lần; khi thêm mô hình ngôn ngữ lớn, tốc độ nhắm mục tiêu nhanh hơn khoảng 5 lần. Thị giác máy tính là công nghệ để máy đọc hình ảnh như mắt người. Đôi mắt đọc hình ảnh làm năng lực tăng 10 lần, còn bộ não đọc chữ làm tốc độ tăng 5 lần. Mắt và não đã được gắn lại với nhau.

Cần nhìn kỹ thêm hai con số này. 10 lần và 5 lần không chỉ có nghĩa là công việc nhanh hơn. Nó giống như việc một người từng xem xét trong một ngày nay được mười người cùng xem, rồi mười người ấy lại thao tác nhanh gấp năm lần. Mục tiêu vì thế đổ tới dồn dập. Vì con người không thể xem xét hết lượng mục tiêu tràn vào, máy lại phải chọn lọc tiếp. Năng lực càng tăng, càng có nhiều việc vượt khỏi bàn tay con người. Càng nhanh, con người càng bị đẩy lùi về phía sau.

Sau khi được đánh khung, mục tiêu ấy được chuyển sang bảng Kanban. Theo giải thích của chuyên trang phân tích Spatial Intelligence, quy trình là như vậy. Bảng Kanban là một bảng chia các cột dọc theo từng công đoạn để chuyển việc qua lại. Như linh kiện trong nhà máy đi từng ô sang công đoạn tiếp theo, mục tiêu cũng chuyển từ cột này sang cột khác. Cũng theo giải thích đó, hệ thống tính thời gian, nhiên liệu, đạn dược và khoảng cách, rồi phân bổ phương tiện phù hợp cho mục tiêu ấy. Máy tính xem máy bay nào ở gần và còn đủ nhiên liệu, vũ khí nào hợp với khoảng cách đó. Rồi ngay trước khi thực hiện, con người phê chuẩn. Mục tiêu được đánh khung, chuyển cột, gắn phương tiện, cuối cùng con người đóng dấu. Dòng chảy ấy không khác băng chuyền trong nhà máy. Chỉ có điều thứ nằm ở cuối dây chuyền không phải linh kiện, mà là mục tiêu.

Một quan chức cấp cao của Bộ Quốc phòng cũng nói đúng dòng chảy ấy. Theo DefenseScoop, Thứ trưởng Bộ Quốc phòng Feinberg nói rằng nhà phân tích làm "từ nhận diện mục tiêu, lập phương án hành động, cho đến tấn công mục tiêu trong một hệ thống". Trong một hệ thống. Trước đây, người tìm mục tiêu, người nghĩ cách đánh và người thực sự đánh là những người khác nhau. Nay cả ba cùng tụ lại trước một màn hình. Hành lang đã ngắn lại. Trên hành lang ngắn ấy, mô hình đi cùng từ đầu đến cuối.

Mô hình sống ở đâu, ba tầng của hạ tầng mật

Muốn đọc thông tin, mô hình phải sống ở một nơi nào đó. Nghĩa là nó phải chạy trên một máy tính nào đó. Khi chúng ta nói chuyện với chatbot, chatbot ấy chạy trong những máy tính lớn của một công ty nào đó. Mô hình không nằm trong điện thoại cầm tay của chúng ta. Trong quân đội cũng vậy. Mô hình sống trong một máy tính ở đâu đó, còn màn hình trong phòng tác

chiến hỏi máy tính ấy và nhận câu trả lời.

Nhưng thông tin của hoạt động quân sự không thể đưa lên bất kỳ máy tính nào. Vì nếu bí mật rò rỉ, con người có thể chết. Vì vậy, tùy theo cấp độ bảo mật của thông tin, ngôi nhà nơi mô hình sống cũng khác nhau. Ngay cả cùng là Claude, tùy loại thông tin nó xử lý, nó sẽ đi vào một ngôi nhà sâu hơn. Có ba tầng. Càng đi từ dưới lên trên, mức khóa càng sâu. Tên các cấp độ sẽ nối nhau xuất hiện, nhưng không cần sợ. Có thể hình dung là tầng hầm 1, tầng hầm 2, tầng hầm 3. Càng đi xuống, cửa càng dày, và số người được ra vào càng ít.

Tầng thứ nhất là đám mây chính phủ thương mại. Mô hình được đặt trên các dịch vụ được quản lý mà các công ty đám mây Mỹ bán cho chính phủ. Theo giải thích của Anthropic, Claude chạy trong môi trường đạt FedRAMP High và các mức IL4, IL5 của Bộ Quốc phòng. Tên các mức này nghe phức tạp, nhưng có thể hiểu là những ngăn xử lý thông tin ngày càng nhạy cảm. Anthropic cho biết môi trường này có cấu trúc 'zero operator access'. Nghĩa là người vận hành không thể can thiệp. Ngay cả nhân viên Anthropic cũng không thể chạm vào hạ tầng suy luận đó. Mô hình vẫn chạy ở đó, nhưng chạy ở một nơi bàn tay của công ty tạo ra nó không với tới.

Cần nhìn kỹ hơn vào cụm từ 'zero operator access' này. Thông thường, khi chúng ta dùng chatbot, công ty vận hành có thể xem chúng ta đã hỏi gì. Nhờ vậy công ty sửa và tinh chỉnh mô hình. Nhưng mô hình được đưa vào mạng của quân đội thì khác. Họ khóa lại để ngay cả công ty tạo ra nó cũng không thể nhìn vào bên trong. Đó là thiết bị nhằm ngăn bí mật quân sự chảy về công ty. Nhưng thiết bị này có một mặt khác. Nếu công ty tạo ra mô hình không thể nhìn vào mô hình, thì công ty cũng không biết mô hình đó đã chỉ ra điều gì bên trong mạng quân sự. Nếu không biết, rất khó buộc họ chịu trách nhiệm. Cái khóa dựng lên để giữ bí mật cũng chặn luôn con đường truy trách nhiệm.

Tầng thứ hai là IL6, cấp secret. Theo các bản tin của Business Wire và The Register, Palantir lưu trữ các dòng Claude 3 và 3.5 trong nền tảng AIP của mình bằng SageMaker của Amazon, rồi đưa chúng vào môi trường đã được chứng nhận IL6. Lưu trữ ở đây nghĩa là đặt mô hình lên một hệ máy tính nào đó để chạy. Có thể hình dung Claude thuê chỗ ở trên máy tính của Amazon, giống như thuê nhà. IL6 là ngăn xử lý được đến cấp 'secret'. Đó là một vị trí bị khóa sâu hơn tầng thứ nhất một nấc. Điều cần nhìn ở đây là công ty tạo ra mô hình là Anthropic, nhưng việc đưa mô hình đó vào mạng quân sự lại do Palantir thực hiện. Anthropic nặn ra bộ não, Palantir đưa bộ não ấy vào căn phòng bí mật. Bộ não nhìn thấy gì và chỉ ra gì trong căn phòng đó diễn ra ở nơi bàn tay của công ty tạo ra bộ não không với tới.

Tầng thứ ba là mạng cách ly không khí, tức air gap, cấp tối mật. Đây là ngăn sâu nhất trong ba tầng. Theo thông báo của Amazon Web Services (AWS), Claude 3.5 Sonnet đã được đưa lên Bedrock, đám mây tối mật của Amazon. Theo DefenseScoop, GPT-4o của đối thủ đã được phê duyệt trên đám mây tối mật chính phủ Microsoft Azure. Đúng như tên gọi air gap, mạng này tách khỏi internet bên ngoài về mặt vật lý như có một khoảng không khí ở giữa. Dây nối bị cắt

hẳn. Bên ngoài không thể đi vào, bên trong cũng không thể rò rỉ ra ngoài.

Ở đây còn có một dòng Anthropic ghi chính thức. Mô hình dành riêng cho chính phủ mang tên Claude Gov đã được chỉnh hành vi để 'refuse less' đối với tài liệu mật. Thông thường, Claude từ chối trả lời các câu hỏi nguy hiểm hoặc nhạy cảm. Nếu hỏi cách chế tạo bom, nó không trả lời. Với câu hỏi về vũ khí, nó cũng dè chừng. Đó là thiết bị an toàn. Nhưng ở nơi xử lý nhiệm vụ mật, sự từ chối ấy lại chặn công việc. Nếu một nhà phân tích quân sự hỏi về vũ khí của đối phương mà mô hình từ chối trả lời, phân tích sẽ dừng lại. Vì vậy công ty cho biết họ đã cố ý hạ ngưỡng đó.

Không thể lướt nhẹ qua dòng này. Thiết bị an toàn vốn được đặt ra để ngăn mô hình giúp vào việc nguy hiểm. Nhưng trong nhiệm vụ quân sự, chính thiết bị an toàn ấy lại trở thành vật cản. Vì vậy công ty đã nới thiết bị đó xuống một nấc trong mô hình dành riêng cho quân đội. Ở nơi an toàn và nhiệm vụ va vào nhau, họ đã bước một bước về phía nhiệm vụ. Đây là điều công ty tự ghi, nên cần ghi lại. Không phải ai đó tiết lộ, mà chính công ty tạo ra mô hình đã tự tay viết trong tài liệu chính thức. Công ty ghi trong chính sách rằng không được dùng cho vũ khí, đồng thời trong cùng giai đoạn lại hạ ngưỡng từ chối của mô hình dành riêng cho quân đội. Hai câu này cùng đúng.

Còn một cửa bên nữa. Đó là cửa bên mang tên đa phương thức. Đa phương thức nghĩa là mô hình không chỉ đọc chữ, mà còn đọc cả hình ảnh và video. Mô hình trước đây chỉ đọc chữ. Mô hình hiện nay, khi được cho xem một bức ảnh, sẽ diễn giải bằng chữ xem bên trong có gì. Theo thông báo đăng trên Business Wire, công ty ảnh vệ tinh Planet đã ký hợp tác dùng suy luận đa phương thức của Claude để biến ảnh vệ tinh được chụp hằng ngày thành 'insight có thể hành động'. Nghĩa là biến những bức ảnh nhìn từ trên cao xuống thành phán đoán chỉ ra phải làm gì.

Vệ tinh chụp Trái Đất mỗi ngày. Những bức ảnh ấy chất thành núi. Mắt người không thể xem hết. Vì vậy mô hình xem trước. Mô hình chỉ ra nơi đã khác hôm qua, dấu vết mới xuất hiện, xe cộ đã di chuyển, rồi viết thành chữ. Đó là phân tích tình báo. Nhưng mô hình làm phân tích ấy lại chính là Claude mà chúng ta dùng để sửa câu chữ nơi làm việc. Đây là bề mặt thương mại nơi một mô hình đa dụng mà ai cũng dùng bước qua ngưỡng phân tích tình báo. Không phải một mô hình được làm riêng cho quân sự, mà là một mô hình bán trên thị trường đã đứng vào vị trí đó. Một mô hình gắn chính sách không được dùng cho vũ khí lại ngồi vào chỗ đọc mặt đất từ vệ tinh và biến nó thành phán đoán. Phán đoán ấy chảy đi đâu, sau khi đi qua cửa bên này, sẽ khó hỏi hơn.

Bốn mô hình rẽ theo bốn hướng, cùng một công nghệ nhưng khác đường đi

Đến đây là câu chuyện về một mô hình Claude. Nhưng Claude không phải là mô hình duy nhất có thể xuất hiện trên màn hình của phòng tác chiến.

Theo Breaking Defense, Văn phòng Kỹ thuật số và Trí tuệ nhân tạo trưởng của Bộ Quốc phòng Mỹ, CDAO, đã trao cho bốn công ty các hợp đồng, mỗi hợp đồng có trần 200 triệu USD. Đó là Anthropic, Google, OpenAI và xAI của Elon Musk. Mỗi công ty 200 triệu USD. Những công ty cho đến hôm qua còn tranh giành khách hàng với nhau, nay được gọi vào cùng một căn phòng của cùng một quân đội. Cũng theo bản tin này, nhiệm vụ cốt lõi của hợp đồng được ghi rõ là 'AI dạng tác tử (agentic AI)'. Dạng tác tử là cách mô hình tự gọi công cụ và nối nhiều bước với nhau. Nghĩa là dù con người không ra lệnh từng bước, mô hình vẫn tự chuyển sang ô tiếp theo. Đây chính là điều Bộ Quốc phòng Mỹ chỉ đích danh yêu cầu khi trả tiền cho bốn công ty. Một mô hình ít cần bàn tay con người hơn. Sức nặng của câu này sẽ được nhìn lại ở cuối mục này.

Bốn công ty mang cùng một công nghệ bước vào cùng một quân đội. Nhưng mỗi công ty lại hiện ra với một dáng vẻ khác.

OpenAI đưa ra 'OpenAI for Government' và đặt ChatGPT Gov dành cho chính phủ lên trên Azure cấp IL5. ChatGPT là chatbot được dùng rộng rãi nhất thế giới. Phiên bản chính phủ của chatbot ấy đã được đặt lên mạng của quân đội. Theo công bố của OpenAI và bản tin của Defense News, OpenAI đã đưa GPT-4 và o1 lên Lattice, nền tảng của công ty phòng thủ drone Anduril, để đánh giá mối đe dọa từ drone của đối phương. Mô hình từng viết văn bản nay chuyển sang vị trí ước lượng mối đe dọa đang tiến đến từ bầu trời. Mô hình từng giúp người ta sửa email cho gọn gàng và mô hình đánh giá drone của đối phương có cùng một dòng máu.

Google đưa ra 'Gemini for Government'. Theo công bố của Google Cloud và bản tin của DefenseScoop, mô hình này trở thành mô hình đầu tiên được đưa lên GenAI.mil, nền tảng trí tuệ nhân tạo cấp toàn Bộ Quốc phòng Mỹ. Cụm từ cấp toàn Bộ có nghĩa là toàn bộ Bộ Quốc phòng cùng dùng. Không phải một đơn vị, mà cả tổ chức khổng lồ mang tên Bộ Quốc phòng dùng trí tuệ nhân tạo trên một nền tảng chung. Mô hình đầu tiên được đặt lên nền tảng ấy là Gemini. Nó chạy trên IL5 trong môi trường tách biệt có tên Google Distributed Cloud, và có lộ trình lên IL6 vào năm 2026, tức cấp bí mật. Hiện nó đã tới trước cửa tầng hầm thứ hai, rồi sẽ đi xuống sâu hơn nữa. Distributed Cloud là một đám mây chủ quyền, nơi dữ liệu được tách khỏi bên ngoài và dữ liệu ấy không được dùng để huấn luyện các mô hình công khai. Sovereign có nghĩa là chủ quyền. Tức là chủ nhân của dữ liệu nắm trọn dữ liệu của mình. Cũng theo công bố đó, Google gắn thêm 'Agent Designer', cho phép cả người không biết viết mã cũng có thể tạo tác tử chỉ bằng ngôn ngữ tự nhiên, và giảm ảo giác bằng search grounding. Search grounding là cơ chế buộc lời mô hình tạo ra phải gắn với sự thật tìm được qua tìm kiếm, qua đó chặn bớt lời sai. Theo công bố của Google, nền tảng này vượt 1 triệu người dùng chỉ sau một tháng. Tuy vậy, con số này chỉ giới hạn trong môi trường không mật.

'Grok for Government' của xAI có màu sắc khác. Theo Cơ quan Dịch vụ Tổng hợp Mỹ, GSA, và công bố của xAI, mô hình kết hợp Grok 4 với deep search và khả năng dùng công cụ này được

cung cấp thông qua hợp đồng của GSA với mức giá 0,42 USD cho mỗi cơ quan trong 18 tháng. 0,42 USD cho 18 tháng.

Gần như miễn phí.

Cùng thông báo đó nêu rõ rằng mô hình này sẽ được triển khai trong các môi trường mật và hạn chế. Đây là một mô hình gần như miễn phí, lại đi vào môi trường mật. Nhưng hợp đồng này có một vùng tối. Theo NPR, chỉ vài ngày trước khi hợp đồng được ký, Grok đã tuôn ra các phát ngôn bài Do Thái trong khoảng 16 giờ và tự gọi mình là "MechaHitler". 16 giờ. Suốt hơn nửa ngày, một mô hình đã phun lời thù hận ra với thế giới. xAI giải thích nguyên nhân là do "một cập nhật ngoài ý muốn trong đường dẫn mã cấp cao". Nghĩa là chỉ cần chạm sai một dòng mã, mô hình đã mất kiểm soát.

Lời giải thích này lại càng nặng nề. Nếu chỉ một dòng mã khiến mô hình mất kiểm soát suốt 16 giờ, điều đó có nghĩa là mô hình ấy có thể bị lay chuyển đến vậy chỉ bởi một đầu ngón tay của người điều khiển. Nếu nó dao động trong cửa sổ trò chuyện, người ta cau mày rồi thôi. Nhưng nếu nó dao động khi đang đánh giá mục tiêu trong môi trường mật, cuối con đường đó là gì. Một mô hình có độ an toàn lung lay đã ký hợp đồng với quân đội chỉ vài ngày sau đó. Mức giá gần như miễn phí và vụ mất kiểm soát vài ngày trước cùng nằm trong một hợp đồng. Có phải nó được đưa vào vì giá rẻ hay không, con dấu có được đóng trước khi vụ mất kiểm soát kịp bị quên đi hay không. Tôi sẽ không khẳng định. Tôi chỉ đặt hai sự việc cạnh nhau. Rằng chỉ trong vài ngày, vụ mất kiểm soát và hợp đồng đã cùng xảy ra.

Bốn công ty cũng tách nhau ở chính sách. Dù mang cùng một công nghệ vào cùng một quân đội, mỗi công ty lại đặt một chốt khóa khác nhau ở cánh cửa đi vào. Theo Defense One và tài liệu của Congressional Research Service (CRS), Bộ Quốc phòng đưa ra tiêu chuẩn "mọi mục đích sử dụng hợp pháp (all lawful use)" đối với bên ký hợp đồng. Nghĩa là miễn hợp pháp thì dùng vào việc gì cũng không bị ngăn. Miễn luật cho phép, dù dùng cho vũ khí tự động hay giám sát, công ty cũng không được phản đối.

Trước tiêu chuẩn này, bốn công ty tách đường. xAI và Google chấp nhận tiêu chuẩn đó, không đặt ranh giới riêng đối với vũ khí tự động hay giám sát đại trà. Vị trí của họ là: nếu luật không cấm, chúng tôi cũng không cấm. OpenAI gắn thêm điều kiện. Và chỉ Anthropic là nêu rõ hai điều từ chối. Giám sát đại trà nhằm vào công dân Mỹ, và vũ khí hoàn toàn tự động. Họ khẳng định sẽ không để mô hình của mình được dùng cho hai việc đó. Vũ khí hoàn toàn tự động là loại vũ khí trong đó máy tự quyết định giết người mà không cần sự phê chuẩn của con người. Ở dòng chảy Maven đã nói ở trên, con người còn đóng con dấu cuối cùng; vũ khí hoàn toàn tự động là thứ xóa luôn con dấu ấy. Anthropic kẻ ranh giới rằng con dấu cuối cùng đó phải được giữ lại cho con người.

Cần nói rõ việc nhìn thêm hai dòng ấy có nghĩa là gì. Công ty chấp nhận "mọi mục đích sử dụng hợp pháp" đã nói rằng, miễn luật không cấm, mô hình của họ có thể được dùng ở bất cứ

đâu. Dùng cho vũ khí tự động cũng được, dùng cho giám sát đại trà cũng được; nếu luật cho phép thì họ không ngăn. Anthropic vẽ thêm hai đường ranh của riêng mình lên trên đó. Dù luật cho phép, chúng tôi vẫn từ chối riêng hai việc này. Giám sát đại trà nhằm vào công dân Mỹ, và vũ khí hoàn toàn tự động tự giết người mà không có sự phê chuẩn của con người. Cùng đi vào một quân đội, nhưng mỗi công ty lại gắn một chốt khóa khác nhau ở cánh cửa đi vào.

Dario Amodei của Anthropic đã dùng lời sắc lạnh khi nói về cách làm của các đối thủ. Theo tài liệu của Congressional Research Service, ông gọi đó là "20 phần trăm là thật, 80 phần trăm là kịch an toàn (safety theater)". Kịch an toàn là lời phê phán rằng họ chỉ giả vờ giữ an toàn. Trên sân khấu thì diễn vai an toàn, nhưng thực chất chỉ là những lời hứa rỗng. Dù cầm cùng một công nghệ, một công ty lại gọi sự an toàn của công ty khác là một vở kịch. Nhưng như đã thấy ở mục trước, mô hình của Anthropic, công ty đã kẻ hai đường ranh ấy, cũng hiện trên màn hình của phòng tác chiến nơi chiến dịch Maduro đang vận hành. Công ty kẻ ranh giới và công ty không kẻ ranh giới, rốt cuộc cùng lấp đầy một màn hình.

Rủi ro mới mang tên tác nhân và ảo giác

Tôi đã tạm gác lại cụm từ AI dạng tác nhân. Bây giờ phải nhìn vào sức nặng của cụm từ ấy.

AI dạng tác nhân là kiểu AI trong đó mô hình tự gọi công cụ và tự thực hiện nhiều bước. Con người không cần kéo nó đi từng ô một; mô hình tự biết chuyển sang ô tiếp theo.

Hãy chuyển điều này thành một hình ảnh dễ hiểu. Mô hình trước đây chỉ làm đúng một việc được giao. Nếu ta nói: "Hãy tìm chiếc xe tải trong video này", nó tìm chiếc xe tải rồi dừng lại. Việc tiếp theo phải do con người ra lệnh. Mô hình dạng tác nhân thì khác. Nếu ta nói: "Hãy tìm điểm đáng ngờ trong video này", nó tự rà soát video, tìm chiếc xe tải, lấy lịch sử di chuyển trước đó của chiếc xe tải ấy từ cơ sở dữ liệu khác, tìm các trường hợp tương tự, rồi gom kết quả lại thành báo cáo. Con người chỉ nói một lần, mô hình tự đi qua nhiều ô và mang câu trả lời về. Nhanh. Giảm việc tay chân. Nhưng càng nhanh, những điểm con người có thể chen vào càng ít. Nếu mô hình nhận nhầm xe tải, trong lúc nó kéo lịch sử di chuyển của chiếc xe tải nhầm đó về và viết cả báo cáo, con người không có chỗ để chen vào mà nói: "Khoan đã, không phải chiếc xe tải đó" rồi dừng nó lại. Ô này nối sang ô kia mà không có con người ở giữa. Con người chỉ có mặt một lần lúc đầu và một lần lúc cuối. Ở giữa, mô hình tự đi một mình.

Ở đây lại chồng thêm một rắc rối cũ: ảo giác. Mô hình ngôn ngữ lớn đôi khi bịa ra những sự thật không tồn tại nhưng nghe rất có vẻ đúng. Người ta gọi đó là ảo giác. Thay vì nói rằng mình không biết, mô hình lấp chỗ không biết bằng những lời nghe hợp lý. Con người trong phòng thi, khi không biết đáp án, cũng có thể nói vòng vo cho có vẻ hợp lý. Mô hình cũng làm như vậy. Chỉ khác là mô hình nói vòng vo với vẻ rất tự tin. Nó đưa ra câu trả lời sai bằng đúng giọng điệu như khi đưa ra câu trả lời đúng. Vì vậy con người khó nhận ra đó là điều được bịa ra.

Search grounding và ontology đã nói ở trên là những cơ chế nhằm ép ảo giác xuống. Chúng buộc mô hình vào sự kiện đã được tìm thấy, buộc vào các đối tượng đã được phân loại, để mô hình không tự ý bịa ra. Trước khi mô hình mở miệng, người ta đặt căn cứ vào tay nó. Ý là: đừng nói vòng vo, chỉ nói những gì ghi trên tấm thẻ này.

Nhưng ép xuống không có nghĩa là xóa bỏ. Dù dựng lan can, con đường vẫn trơn. Ba thứ cùng chảy vào pipeline thu hẹp mục tiêu: ảo giác của mô hình, lỗi tìm kiếm, và thiên lệch dữ liệu. Mô hình có thể bịa ra điều không có. Kết quả tìm kiếm được kéo về có thể sai. Dữ liệu được gom từ đầu có thể đã nghiêng về một phía. Ba thứ này đi vào một đầu của pipeline, rồi đi ra ở đầu kia dưới dạng mục tiêu.

Lỗi đó không dừng lại như một sai sót trên bàn phím. Nó hiện ra bằng cái chết của con người. Câu này là câu nặng nhất trong mục này. Trong cửa sổ chat, nếu mô hình trả lời sai, người dùng có thể cười cho qua. Hỏi lại là được. Nhưng trên màn hình trong phòng tác chiến, nếu mô hình chỉ nhầm đối tượng, ở đầu bên kia không có ai để hỏi lại. Ở đầu bên kia là một mục tiêu bị ngắm nhầm. Cùng là ảo giác, nhưng trong cửa sổ chat nó có thể kết thúc như một chuyện đùa; trong phòng tác chiến nó thành điểm cuối của một đời người. Vẫn là cùng một ảo giác, chỉ khác nơi nó rơi xuống. Trong pipeline nơi các điểm con người can thiệp bị giảm bởi mô hình dạng tác nhân, và trách nhiệm bị phân tán vì các phần có thể thay thế cho nhau, lỗi này không được ghi lại như sai lầm của một ai, mà chỉ còn lại dưới dạng cái chết của một con người.

Có một con số do một cơ quan truyền thông đưa ra. Theo bản tin của WION, một cơ quan truyền thông Mỹ, Bộ Tư lệnh Trung tâm Hoa Kỳ (CENTCOM) đã dùng hệ thống tích hợp Claude và Maven để tạo ra khoảng 1.000 ứng viên mục tiêu ưu tiên kèm dữ liệu tác chiến trong vòng 24 giờ. 1.000 mục tiêu trong 24 giờ.

Một nghìn trong một ngày. Hơn bốn mươi trong một giờ. Để xem xét một mục tiêu cho đúng, con người phải gom tài liệu, lục lại hồ sơ cũ, kiểm tra xem xung quanh có thường dân hay không. Nếu phải làm việc đó hơn bốn mươi lần trong một giờ, bàn tay con người không với tới được. Vì vậy máy móc thu hẹp trước.

Tuy nhiên, con số này là một thông tin báo chí chưa được kiểm chứng. Tôi chỉ ghi lại như một tuyên bố của truyền thông. Nếu đúng, ý nghĩa của nó rất nặng. Trước khi con người xem xét từng ứng viên một, máy móc đã thu hẹp danh sách xuống còn 1.000. Việc thẩm tra của con người bắt đầu trên 1.000 cái tên đã được lọc sẵn ấy. Con người nhìn vào 1.000 cái tên đó, rồi đóng dấu chấp thuận hoặc trả lại. Nhưng người chọn ra 1.000 cái tên ấy không phải là con người.

Ở đây chồng lên cái bóng của Lavender từng thấy ở Gaza. Ở đó, danh sách cũng do máy lập ra, còn con người đóng dấu trên danh sách ấy, mỗi người 20 giây. Người không có tên trong danh sách thậm chí không được xem xét; người bị đưa nhầm vào danh sách nhận dấu chấp thuận mà không biết rằng mình đã bị máy chọn. Con số 1.000 ở Venezuela cũng đứng ở cùng vị

trí ấy. Những gì không lọt vào 1.000 đó không chạm tới cả ngưỡng cửa thẩm tra; những gì bị đưa nhầm vào 1.000 đó đi đến trước mặt con người với sức nặng của việc đã được máy chọn. Sự thật rằng con người đóng dấu ở bước cuối không xóa được sự thật rằng máy đã chọn ứng viên ở bước đầu. Con dấu cuối cùng chỉ được đóng trên chiếc bàn mà màn hình đã dọn sẵn. Cái gì được đặt lên bàn, màn hình đã quyết định trước.

Trở lại chỗ trống

Bây giờ trở lại câu hỏi ban đầu. Sức sát thương đến từ mô hình, hay đến từ bàn tay đưa mô hình lên chiến trường?

Công ty tạo ra mô hình nói: mô hình của chúng tôi không phải là vũ khí. Về cấu trúc, hành vi tấn công đã bị chặn, và theo chính sách, việc dùng cho vũ khí bị cấm. Công ty đưa mô hình lên chiến trường nói: chúng tôi chỉ cài đặt công cụ, còn người bóp cò là con người. Lực lượng thực hiện chiến dịch nói: máy chỉ khuyến nghị, còn quyết định là do con người đưa ra.

Cả ba câu đều đúng. Vì thế mới thành vấn đề.

Hãy nhìn lại con đường đã đi qua trong mục này. Mô hình là bộ não đọc và viết ngôn ngữ. Tự nó không có cò súng. Tích hợp là việc mở đường bằng ontology để bộ não ấy đọc thông tin chiến trường, rồi vẽ các hộp trên màn hình của Maven. Tự nó không nhắm vào con người. Vận hành là việc đóng dấu lên những chiếc hộp ấy. Tự nó chỉ chuyển động trên những gì màn hình đã dọn ra. Ba vị trí ấy mỗi bên chỉ nhìn một ô trước mặt mình. Bộ não nói mình chỉ đọc thông tin, màn hình nói mình chỉ sắp xếp, bàn tay nói mình chỉ chọn.

Sức sát thương không nằm trọn trong mô hình, cũng không nằm trọn trong tích hợp hay vận hành. Nó mắc ở khoảng hở giữa ba vị trí ấy. Trách nhiệm rò rỉ qua khoảng hở đó. Mô hình đọc thông tin, tích hợp gọt thông tin ấy thành hình dạng của mục tiêu, vận hành đóng dấu lên mục tiêu đó. Mỗi bên chỉ chịu trách nhiệm cho ô của mình. Giữa ô này và ô kia, vị trí chịu trách nhiệm cho toàn bộ quá trình tạo ra cái chết của một con người vẫn để trống.

Một phát biểu của một nhân vật cấp cao trong Bộ Quốc phòng làm khoảng trống ấy rộng thêm. Theo Defense One, ông nói rằng quân đội sẽ không bị trói vào một công ty nào, và rằng họ sẽ "không bao giờ phụ thuộc vào một nguồn duy nhất nữa (never again single-threaded)". Nếu chỉ dựa vào một công ty, chiến dịch sẽ dừng lại khi công ty đó tăng giá, cắt nguồn cung, hoặc do dự. Sự do dự của Anthropic đã nói ở phần trước chính là rủi ro ấy. Vì vậy, quân đội chọn con đường đưa cùng lúc mô hình của nhiều công ty vào mạng mật, rồi thay đổi tùy theo mục đích. Nghe nói có khoảng tám đơn vị. Việc này dùng Claude, việc kia dùng Gemini, việc khác dùng Grok. Nếu một mô hình do dự, họ thay bằng mô hình khác. Mô hình trở thành linh kiện. Một linh kiện có thể tháo ra lắp vào.

Linh kiện thường không có tên rõ ràng. Khi động cơ ô tô ngừng chạy, chúng ta không đổ lỗi cho một con ốc nào đó. Mô hình càng trở thành linh kiện có thể thay thế, cái chết mà linh kiện

ấy chỉ tới thuộc về ai lại càng mờ đi. Claude đã chỉ điểm chăng, Gemini đã chỉ điểm chăng, hay Grok đã chỉ điểm chăng. Ngay cả việc hôm đó mô hình nào hiện trên màn hình cũng không còn rõ trong một hệ thống thay lắp như vậy. Khi có nhiều linh kiện, ngón tay truy trách nhiệm không biết phải chỉ vào đâu. Câu nói không phụ thuộc vào một nguồn duy nhất vừa có nghĩa là không bị trói vào một công ty, vừa có nghĩa là việc buộc một công ty chịu trách nhiệm cũng trở nên khó hơn. Khi sự phụ thuộc bị phân tán, rủi ro cũng phân tán, nhưng trách nhiệm cũng tan theo.

Ở phần trước, Karp ví thời đại này với 'khoảnh khắc Oppenheimer'. Ý ông là, giống như các nhà vật lý đã tạo ra vũ khí hạt nhân, những kỹ sư hôm nay cũng đang cầm trong tay một công cụ mạnh. Nhưng với Oppenheimer, ít nhất có một điều rõ ràng. Đó là ông đã tạo ra thứ gì. Quả bom là một vật thể duy nhất, và bàn tay chế tạo nó, bàn tay thả nó, cái miệng ra lệnh cho việc đó được tách bạch thành ba. Lịch sử có thể ghi lại ba bên ấy riêng rẽ. Bây giờ thì khác. Có nhiều công ty tạo ra bộ não, bộ não ấy được thay lắp, công ty dựng màn hình và bàn tay đóng dấu lại không phải một. Cái gì đã chỉ tới cái chết bị rải ra trong đống linh kiện. Trong khoảnh khắc Oppenheimer, ít nhất còn có một khuôn mặt để chịu trách nhiệm. Trong khoảnh khắc của khoảng trống, ngay cả khuôn mặt ấy cũng bị phân tán.

Tôi sẽ không khép lại câu hỏi. Sức sát thương đến từ mô hình hay từ bàn tay, phần này chưa ấn định câu trả lời. Vì không thể ấn định. Nếu tách riêng mô hình, nó chỉ là một bộ não viết chữ. Nếu tách riêng tích hợp, nó chỉ là một màn hình sắp xếp thông tin. Nếu tách riêng vận hành, nó chỉ là bàn tay đóng dấu. Trong ba thứ ấy, không thứ nào tự nó giết người. Nhưng khi ba thứ gặp nhau, con người chết. Sức sát thương nằm vắt qua cả ba, vì thế trách nhiệm cũng rò rỉ qua khe giữa ba thứ ấy.

Đó là nơi phần này đi tới. Thời cũ, khi người ta chế tạo vũ khí, vũ khí là một vật thể. Có nhà máy làm súng, có người lính cầm súng, có người ra lệnh bắn. Ba bên rất rõ. Vũ khí hôm nay không phải một vật thể, mà là một sợi dây được bện qua bàn tay của nhiều công ty. Công ty nặn ra bộ não, công ty dựng màn hình, công ty đưa nó lên mạng, quân đội quyết định thay lắp. Khi muốn hỏi nút thắt nào trên sợi dây ấy đã chỉ tới cái chết, mỗi nút thắt lại nằm trong một bàn tay khác. Mỗi bàn tay chỉ chịu trách nhiệm về nút thắt mình cầm. Không có bàn tay nào nắm cả sợi dây.

Trách nhiệm bị rò rỉ ấy đi về đâu, đó là khoảng trống mà chương sau sẽ lần theo. Đèn ở Caracas đã sáng trở lại, và Maduro đã đứng trước tòa. Nhưng trước sợi dây đã đưa ông tới tòa án ấy, sợi dây được bện bởi bộ não, màn hình và bàn tay, vẫn chưa có ai đứng ra.

# Phần 3 Khoảng trống Chương 5 Hộp đen và trách nhiệm đang bốc hơi

## 1. Thiên kiến tự động hóa: tâm lý khiến con người tin vào máy móc

Hãy quay lại chiếc bàn làm việc ấy.

Trước mặt một sĩ quan tình báo ở Gaza có một màn hình. Trên đó hiện lên một cái tên. Bên cạnh cái tên là một con số. Từ 1 đến 100. Đó là điểm do máy chấm. Viên sĩ quan nhìn tên, nhìn điểm, rồi chỉ kiểm tra một điều. Có phải nam giới không. Nếu đúng, ông ta chuyển sang người tiếp theo. Thời gian ông ta dành cho một mục tiêu là 20 giây.

Chúng ta đã thấy trong 20 giây ấy chuyện gì xảy ra. Chúng ta cũng đã thấy rằng ông ta không xem xét tội trạng của mục tiêu. Điều ông ta kiểm tra là giới tính. Vì sao Lavender chọn người đó, mẫu liên lạc của người đó có thật sự giống mẫu của lực lượng vũ trang hay không, liệu đó có thể là một người khác trùng tên hay không, ông ta không có thời gian để nhìn vào những việc như vậy. 20 giây không phải là thời gian thẩm tra. Đó là thời gian phê duyệt.

Đổ lỗi cho sự lười biếng của một sĩ quan trong cảnh này thì dễ. Nhưng như vậy là đọc sai cảnh ấy. Đặt bất kỳ ai vào cùng vị trí đó, chuyện tương tự cũng sẽ xảy ra. Khi màn hình đưa ra câu trả lời, thời gian chỉ có 20 giây, cấp trên thúc phải xử lý nhanh, con người gần như luôn chọn cùng một cách. Máy nói vậy thì nhận là vậy. Lựa chọn này có tên gọi. Đó là thiên kiến tự động hóa.

Tâm lý này không chỉ có trên chiến trường. Chúng ta cũng gặp nó mỗi ngày. Khi hệ thống dẫn đường trên ô tô chỉ đường, chúng ta không xét xem con đường ấy có thật sự nhanh nhất hay không, mà xoay vô lăng theo hướng đó. Khi máy tính đưa ra con số, chúng ta không tính lại để kiểm tra. Khi phần mềm kiểm tra chính tả gạch đỏ một từ, chúng ta sửa từ ấy mà không mở từ điển xem nó có thật sự sai không. Việc chấp nhận câu trả lời do máy đưa ra thường là hợp lý. Máy thường đúng hơn chúng ta, và nghi ngờ từng việc một thì mệt. Vì thế, thiên kiến tự động hóa không phải là sản phẩm của sự ngu dốt, mà là kết quả tự nhiên của một tâm trí chạy theo hiệu suất. Vấn đề nằm ở chỗ tâm trí ấy hoạt động như nhau trước màn hình dẫn đường và trước màn hình chọn mục tiêu. Chỉ khác là ở một bên, ta đi sai đường thêm 5 phút; ở bên kia, một người chết.

Hai nhà tâm lý học là những người đầu tiên định nghĩa rõ ràng điều này. Kathleen Mosier và Linda Skitka. Vào cuối thập niên 1990, hai người nghiên cứu chuyện gì xảy ra khi con người làm việc cùng các thiết bị hỗ trợ tự động. Điều họ tìm thấy là như sau. Khi con người dựa vào câu trả lời do máy đưa ra, họ có xu hướng bỏ qua cả thông tin đúng đang ở ngay trước mắt nếu thông tin ấy trái với câu trả lời đó. Thay vì tự xem xét và tính lại, họ chọn con đường chấp nhận nguyên kết luận của máy. Mosier và Skitka gọi đây là 'con đường ít tốn sức nhận thức nhất'. Nghĩa là tâm trí trôi về phía phải dùng ít đầu óc hơn.

Thiên kiến này có hai gương mặt. Một là sai lầm do bỏ sót. Vì máy không phát cảnh báo, con người tưởng không có chuyện gì và buông tay. Gương mặt kia là sai lầm do đi theo. Máy đưa ra tín hiệu sai, nhưng con người vẫn thực hiện nguyên như vậy. 20 giây ở Gaza gần với gương mặt thứ hai hơn. Máy hiện lên một cái tên, và con người đi theo cái tên ấy.

Thiên kiến tự động hóa có một cặp đối diện như trong gương. Đó là né tránh thuật toán (algorithm aversion). Nếu thiên kiến tự động hóa là tin máy móc quá mức, thì né tránh thuật toán là nghi ngờ máy móc quá sớm. Đó là việc máy đã đưa ra câu trả lời đúng, nhưng con người vẫn khăng khăng giữ phán đoán của mình. Cả hai đều nguy hiểm. Nhưng mức nguy hiểm của chúng không giống nhau. Vì sai lỗi của trí tuệ nhân tạo hiện nay thường không hỏng lộ liễu, mà sai theo cách trông có vẻ đúng. Câu trả lời hiện trên màn hình trông như thật nhưng thực ra lại sai, đó là loại lỗi khó xử lý nhất. Vì vậy người ta lo về thiên kiến nhiều hơn là né tránh.

Có một ví dụ điển hình về bi kịch do né tránh gây ra. Năm 1988, tuần dương hạm USS Vincennes của Hải quân Mỹ đã bắn hạ chuyến bay Iran Air 655 trên Vịnh Ba Tư. Đó là một máy bay chở khách dân sự. Toàn bộ 290 hành khách và thành viên phi hành đoàn đều thiệt mạng. Hôm đó, hệ thống tác chiến Aegis của tàu đã nhận diện đúng chiếc máy bay ấy là máy bay chở khách dân sự. Nhưng các thủy thủ đang ở trong trạng thái căng thẳng cực độ đã không tin đầu ra đó. Điều máy nói là đúng, con người lại đảo ngược thành sai. Trực giác sai lầm của con người đã phủ lên phán đoán đúng của máy. Đây là khuôn mặt của né tránh thuật toán. Nếu Gaza là việc xảy ra vì con người tin máy quá nhiều, thì USS Vincennes là việc xảy ra vì con người không tin máy đủ.

Đặt hai sự kiện cạnh nhau, một điều trở nên rõ ràng. Câu trả lời đúng không phải là tin máy vô điều kiện, cũng không phải là nghi ngờ máy vô điều kiện. Việc còn lại thật sự của con người là đo mức đáng tin của máy trong từng hoàn cảnh và chỉnh kích thước niềm tin cho đúng. Nhưng để làm được việc đo đó, cần có hai thứ. Thời gian và một cửa sổ để nhìn vào bên trong. Sĩ quan tình báo ở Gaza không có thời gian, và cũng không có cửa sổ để nhìn vào bên trong Lavender.

Lý do thiên kiến tự động hóa đáng sợ là nó không phải căn bệnh chỉ tìm đến người lười biếng. Trái lại, máy càng thường xuyên đoán đúng, thiên kiến ấy càng ăn sâu. Một bài tổng quan về các nghiên cứu y học liên quan đến thiên kiến tự động hóa đã báo cáo một sự thật lạ. Khi được một hệ thống tự động hóa luôn có độ chính xác cao giúp đỡ, con người lại mắc nhiều lỗi hơn so với khi dùng một hệ thống tự động hóa thỉnh thoảng sai. Nghĩa là máy càng đoán đúng, con người càng ngủ sâu hơn. Nếu màn hình đã đúng 99 lần trong 100 lần, người ta sẽ tin rằng lần thứ 100 cũng đúng. Dù lần thứ 100 ấy là tọa độ của một trường học.

Ở đây hãy nhìn vào những con số mà một hệ thống của Mỹ cho thấy. Đó là Maven Smart System. Đây là hệ thống nhận diện mục tiêu lấy tên từ Project Maven, dự án đầu tiên mà Lầu Năm Góc bắt tay vào trí tuệ nhân tạo. Theo một bản tin được các nhà nghiên cứu West Point

trích dẫn, Quân đoàn Dù 18 của Mỹ đã dùng hệ thống này để giảm thời gian cần thiết nhằm xác định mục tiêu chính xác xuống còn 1 phần 100. Mà chỉ với 20 người. Trước đây, để làm cùng công việc ấy cần hơn 2.000 nhân sự hỗ trợ. Công việc của 2.000 người, nay 20 người làm, nhanh hơn 100 lần. Và cùng nghiên cứu đó cho biết Bộ Tư lệnh Trung tâm Mỹ đã dùng rộng rãi hệ thống này trong cuộc chiến với Iran. Việc từng mất vài giờ, vài ngày, nay giảm xuống còn vài phút. Người ta nói hệ thống này là nền tảng giúp Mỹ có thể tấn công hàng nghìn mục tiêu của Iran trong thời gian ngắn.

Khi đặt những con số này lên trên thiên kiến tự động hóa, bức tranh hiện ra rõ hơn. Tốc độ xử lý mục tiêu nhanh hơn 100 lần cũng có nghĩa là thời gian mà một con người có thể dành cho một mục tiêu giảm xuống còn 1 phần 100. Thứ nhanh hơn là niềm tự hào của máy. Thứ bị rút ngắn là thời gian của con người. Và thời gian của con người càng bị rút ngắn, thiên kiến tự động hóa càng ăn sâu. Trong một hệ thống coi tốc độ là đức tính, việc con người dừng lại để chất vấn không còn là đức tính nữa, mà trở thành vật cản. Hệ thống ấy được thiết kế để con người không chất vấn, mà chỉ đóng dấu thật nhanh.

Năm 2024, Ủy ban Chữ thập đỏ Quốc tế đã đưa ra cảnh báo về vấn đề này. Tổ chức này chỉ ra nguy cơ con người rơi vào thiên kiến tự động hóa trước các hệ thống hỗ trợ ra quyết định quân sự, tức trí tuệ nhân tạo gợi ý mục tiêu. Điều Chữ thập đỏ lo ngại không phải là một lỗi kỹ thuật vụn vặt. Đó là trạng thái trong đó con người vẫn ngồi trong chuỗi quyết định, nhưng trên thực tế lại không phán đoán. Chỉ cần nhớ lại câu đã thấy ở trước. Việc con người có mặt trong chuỗi và việc con người thật sự đang phán đoán không phải là một.

Vậy nếu được huấn luyện, con người có thể thắng thiên kiến này không. Có một nghiên cứu đã thật sự kiểm tra câu hỏi ấy. Đó là thí nghiệm với các học viên Học viện Quân sự Mỹ, West Point. Lauren Kahn của Đại học Georgetown, Michael Horowitz của Đại học Pennsylvania, và Laura Resnick Samotin của West Point đã thực hiện vào tháng 5 năm 2025. Họ giao cho 236 học viên một nhiệm vụ nhận diện máy bay. Khi một máy bay xuất hiện trên màn hình, học viên phải nhìn hình dạng cánh và đuôi để xác định đó là máy bay phe mình hay máy bay địch. Sau khi học viên đưa ra phán đoán trước, một lời 'khuyên' mới hiện lên. Có lúc họ được cho biết đó là lời khuyên của thuật toán trí tuệ nhân tạo, có lúc là lời khuyên của một nhà phân tích con người. Học viên có thể nhìn lời khuyên ấy rồi giữ nguyên câu trả lời của mình, hoặc đổi câu trả lời.

Nhóm nghiên cứu đo hai điều. Khi người tham gia làm theo lời khuyên sai, họ gọi đó là thiên kiến tự động hóa. Khi người tham gia bác bỏ lời khuyên đúng, họ gọi đó là né tránh thuật toán. Kết quả đem lại một nửa sự an ủi. Tỷ lệ làm theo lời khuyên sai của các học viên West Point chưa bằng một nửa so với người bình thường. Mẫu người bình thường làm theo lời khuyên sai trong 9% trường hợp, còn các học viên chỉ dừng ở 3,9%. Người đã được giáo dục quân sự và được huấn luyện về trí tuệ nhân tạo tỉnh táo hơn một chút trước câu trả lời của máy. Điều đáng chú ý là các học viên đã phản ứng với điều gì. Họ không bị chi phối nhiều bởi việc lời khuyên đến từ trí

tuệ nhân tạo hay con người, mà nhạy hơn với mức độ hệ thống ấy đã được kiểm chứng. Trước một hệ thống được nói là 'đã được huấn luyện đầy đủ', họ đổi câu trả lời. Trước một hệ thống 'vẫn đang thử nghiệm', họ giữ phán đoán của mình. Nhóm nghiên cứu gọi điều này là 'niềm tin chính đáng'. Đó là thái độ không tin máy một cách mù quáng, cũng không nghi ngờ máy một cách mù quáng, mà điều chỉnh mức độ tin cậy theo việc cỗ máy ấy đáng tin đến đâu.

Nếu chỉ nhìn đến đây thì có vẻ còn hy vọng. Nhưng khi đặt nghiên cứu này trước màn hình của Gaza và Iran, sự an ủi ấy nhanh chóng nguội đi. Vì chiến trường thật có một thứ mà thí nghiệm West Point không có. Thời gian.

Các học viên được cho 7 đến 10 giây để nhận dạng một máy bay. Ngay cả khoảng thời gian ngắn ấy cũng là thời gian trong phòng thí nghiệm. Chiếc máy bay trên màn hình không thả xuống bất cứ thứ gì. Nếu trả lời sai, học viên chỉ mất 0,1% điểm thưởng. Ở đầu ngón tay bấm nhầm, không có ai chết.

Chiến trường thì khác. Sĩ quan tình báo ở Gaza được trao 20 giây cho một mục tiêu. Nếu trong cuộc chiến với Iran phải xử lý 1.000 mục tiêu trong vòng 24 giờ, thời gian con người có thể dành cho mỗi mục tiêu chỉ hơn 1 phút một chút. Và ở cuối 1 phút ấy là bom thật. Trên chiến trường drone ở Ukraine, có tin rằng chuỗi phát hiện và tấn công mục tiêu đã rút xuống còn 18 phút. Thời gian càng bị rút ngắn, khoảng dừng để con người đứng lại và cân nhắc cũng càng hẹp đi.

Thiên kiến tự động hóa càng sâu khi thời gian càng ít. Không có thời gian để cân nhắc, người ta sẽ không cân nhắc nữa. Khi không cân nhắc, câu trả lời của máy trở thành câu trả lời của con người. 'Niềm tin chính đáng' mà các học viên West Point cho thấy là một phẩm chất chỉ vận hành khi còn thời gian để xem xét. Khi thời gian bị tước đi, phẩm chất ấy cũng biến mất theo. Ngay cả người được huấn luyện tốt, trước 20 giây, cũng cử động ngón tay giống người không được huấn luyện tốt. Giữa 7 giây trong phòng thí nghiệm và 20 giây trên chiến trường có một khác biệt nặng đến mức không đo được. Một đầu là điểm số, đầu kia là con người.

Công bằng mà nói, cần nói rõ một điều. Các tác giả của nghiên cứu West Point đã tự ghi lại giới hạn trong phát hiện của họ. Điều họ đo là một gương mặt của thiên kiến tự động hóa, tức sai lầm làm theo thông tin sai. Sai lầm buông tay vì máy không phát cảnh báo thì, do thiết kế thí nghiệm, không thể đo được. Họ cũng viết rằng, đối với cách Israel dùng Lavender, có các phân tích sau này cho rằng sự việc không phải là con người bị buộc phải quyết định trong vài giây như các bản tin ban đầu thuật lại, mà là hệ thống phản ánh mức tin cậy và quy tắc giao chiến do chỉ huy định sẵn. Cuốn sách này không đóng đinh một bản tin nào là sự thật cuối cùng. Nhưng dù chọn cách giải thích nào, câu hỏi còn lại vẫn như nhau. Nếu mức tin cậy do chỉ huy định sẵn vượt quá độ chính xác của máy, thì phê chuẩn trong 20 giây không phải là lỗi cá nhân, mà là kết quả của quy tắc từ cấp trên.

Ngay cả các tác giả của nghiên cứu này cuối cùng cũng thừa nhận một điều. Thiên kiến tự động hóa có thể không phải là vấn đề của cá nhân, mà là vấn đề của tổ chức. Nếu quy trình tác

chiến chuẩn của quân đội ra lệnh rằng 'hãy tin khuyến nghị của trí tuệ nhân tạo vượt quá mức mà độ chính xác của nó bảo đảm', thì trách nhiệm về kết quả không nằm ở cá nhân ngồi trước màn hình. Nó nằm ở những người từ trên ra lệnh làm như vậy. Các sĩ quan tình báo ở Gaza nói rằng họ được chỉ thị chấp nhận khuyến nghị của Lavender nếu chỉ cần xác nhận mục tiêu là nam giới. Dưới chỉ thị ấy, phê chuẩn trong 20 giây không phải là sự yếu đuối của một cá nhân, mà là thiết kế của hệ thống.

Vì vậy, thiên kiến tự động hóa có hai lớp. Một lớp là tâm trí con người. Tâm trí muốn tin vào máy, muốn bớt dùng đầu óc. Lớp còn lại là thủ tục được thiết kế để dùng chính tâm trí ấy. Mệnh lệnh xử lý nhanh, thời hạn 20 giây, chỉ thị không được nghi ngờ. Ở nơi hai lớp này gặp nhau, con người ngồi trong chuỗi vận hành và đặt phán đoán của mình xuống.

Nhưng vẫn còn một câu hỏi khó hơn. Dù viên sĩ quan ấy có muốn nghi ngờ, ông ta có thể nghi ngờ điều gì? Ông ta có biết vì sao Lavender lại cho người đó điểm cao không? Phần tiếp theo sẽ đi vào câu hỏi ấy.

## 2. Thế lưỡng nan hộp đen: cái chết không thể giải thích

Hãy quay lại ngôi trường ở Minab.

Minab, thành phố cảng ở miền nam Iran. Trường tiểu học Sharezeh Tayebeh. Khoảng 10 giờ 20 phút sáng ngày 28 tháng 2 năm 2026. Tiết học đầu tiên vừa kết thúc và tiết thứ hai bắt đầu. Trẻ em ngồi tại chỗ, các cô giáo đứng trước bảng. Vào thời điểm ấy, một quả tên lửa được cho là Tomahawk do Mỹ sản xuất rơi xuống tòa nhà đó. Đây là thông tin do The New York Times và CNN đưa tin. Hai mươi sáu trẻ em từ bảy đến mười hai tuổi và các cô giáo thiệt mạng. Con số thống kê khác nhau, từ 156 đến hơn 175 người.

Tên lửa không bắn trượt. Nó bay chính xác đến tọa độ. Vấn đề nằm ở chính tọa độ ấy. Theo các bản tin, tọa độ đó đã có từ 10 năm trước. Không rõ nơi ấy từng có cơ sở quân sự hay ngay từ đầu đã bị nhập sai. Nhưng có một điều rõ ràng. Vào năm 2026, tại vị trí ấy là một ngôi trường, còn trong dữ liệu của hệ thống lại có một mục tiêu. Hai thứ đó không khớp nhau.

Đến đây, ta muốn hỏi trách nhiệm. Ai đã đặt tọa độ ấy ở đó? Ai đã không cập nhật nó? Vì sao hệ thống lại đề xuất tọa độ ấy làm mục tiêu? Trước những câu hỏi này, chúng ta nhanh chóng đụng vào một bức tường. Tên của bức tường ấy là hộp đen.

Hộp đen nghĩa là một chiếc hộp không nhìn thấy bên trong. Dữ liệu đi vào, kết quả đi ra, nhưng từ bên ngoài không thể thấy điều gì diễn ra ở giữa. Nhiều hệ thống trí tuệ nhân tạo quân sự hiện nay là như vậy. Trong đó có deep learning, tức những hệ thống tự học từ lượng dữ liệu khổng lồ để tìm ra mẫu hình. Những hệ thống này quan sát hàng triệu trường hợp rồi tự tạo ra quy tắc. Quy tắc ấy không phải do con người trực tiếp viết vào. Máy đã kéo nó lên từ trong dữ liệu. Vì thế, khi máy đưa ra một đáp án nào đó, ngay cả người tạo ra máy cũng thường không thể giải thích rõ vì sao nó đưa ra đáp án ấy.

Hãy nghĩ lại về Lavender. Lavender nhìn vào những tín hiệu nhỏ như thói quen liên lạc của một người, người đó có thường đổi điện thoại hay không, có thuộc nhóm trò chuyện nào không, rồi chấm điểm. Nhưng các tín hiệu ấy được cộng với trọng số nào để thành 90 điểm hay 30 điểm, phần tính toán bên trong không hiện lên trên màn hình. Điều sĩ quan tình báo nhìn thấy chỉ là kết quả. Tên và điểm số. Dù ông ta muốn hỏi kỹ căn cứ của điểm số ấy, phần bên trong để hỏi lại không hiện ra trước mắt ông ta. Muốn nghi ngờ cũng không có gì trên màn hình để nghi ngờ. Thiên kiến tự động hóa và hộp đen ăn khớp với nhau như vậy. Con người bắt đầu tin máy, đồng thời cũng không thể nhìn vào bên trong máy. Niềm tin và sự mờ đục nắm tay nhau.

Đây là điểm tách vũ khí cũ khỏi vũ khí trí tuệ nhân tạo. Hãy nghĩ đến một khẩu pháo bắn lệch. Nếu quả đạn rơi nhầm chỗ, chúng ta có thể lần ngược nguyên nhân. Nhắm sai, đo gió sai, hay lượng thuốc súng không đúng. Chuỗi nguyên nhân hiện ra trước mắt. Nhưng nếu muốn lần ngược vì sao Lavender cho một người nào đó 92 điểm, chuỗi ấy biến mất vào hàng triệu kết nối bên trong cỗ máy. Nếu đó là quy tắc do con người viết vào, ta có thể mở quy tắc ấy ra đọc. Còn quy tắc mà máy rút lên từ dữ liệu thì không có tờ giấy nào để mở ra. Nó rải trong một tấm lưới khổng lồ của những con số.

Ta đã thấy ở phần trước rằng Lavender không phân biệt được đời sống thường ngày của con người với hành vi của lực lượng vũ trang. Người thường xuyên đổi điện thoại có thể là thành viên lực lượng vũ trang, cũng có thể chỉ là một người nghèo. Người ở cùng một phòng trò chuyện có thể là đồng sự, cũng có thể chỉ là hàng xóm. Cỗ máy gán điểm cho mẫu hình mà không biết sự khác nhau ấy. Vấn đề nằm ở chỗ máy không tự giải thích được vì sao điểm số lại được chấm như vậy. Màn hình không nói cho ta biết giữa người 90 điểm và người 89 điểm khác nhau ở đâu, một điểm chênh lệch đến từ đâu. Dù chính một điểm ấy đã chia đôi sự sống và cái chết.

Viện Nghiên cứu Hòa bình Quốc tế Stockholm, viết tắt là SIPRI, đã bàn về vấn đề này trong một báo cáo vào tháng 8 năm 2025. Đó là báo cáo phân tích cách thiên lệch của trí tuệ nhân tạo quân sự va chạm với việc tuân thủ luật nhân đạo quốc tế. Bài viết của Laura Bruun và Marta Bo chỉ ra thiên lệch đi vào từ đâu. Trọng tâm là dữ liệu huấn luyện. Trí tuệ nhân tạo giống dữ liệu mà nó đã học. Nếu dữ liệu nghiêng về một phía, phán đoán của máy cũng nghiêng về một phía. SIPRI nói thiên lệch này gây hại theo hai cách. Một là nhận dạng sai. Tức là xác định nhầm người hoặc vật cần được bảo vệ thành mục tiêu. Hai là thất bại trong nhận dạng. Tức là không nhận ra người cần được bảo vệ là người cần được bảo vệ. Trường học ở Minab gần với trường hợp thứ nhất. Trong dữ liệu của hệ thống, tọa độ ấy là mục tiêu, còn thông tin rằng mục tiêu ấy thật ra là một trường học thì không chạm được tới dữ liệu.

SIPRI chia nhỏ hơn nơi thiên lệch đi vào. Nó có thể xuất hiện ở giai đoạn thu thập dữ liệu. Nếu dữ liệu của một nhóm thì dư thừa còn dữ liệu của nhóm khác thì thiếu, máy sẽ nhận ra nhóm có

dữ liệu dư tốt hơn và thường bỏ sót nhóm thiếu dữ liệu. Nó cũng có thể xuất hiện ở giai đoạn thiết kế hệ thống. Trong lựa chọn lấy điều gì làm tín hiệu của mục tiêu, phỏng đoán của con người đã thấm vào. Và thiên lệch ấy không đứng yên một chỗ. Độ nghiêng của dữ liệu huấn luyện lan sang khuyến nghị mục tiêu. Một độ lệch nhỏ ở đầu vào trở thành sai lầm lớn ở đầu ra.

Báo cáo của SIPRI đưa ra một nhận xét nặng ký. Thiên lệch này không thể bị loại bỏ hoàn toàn. Dù chỉnh sửa dữ liệu đến đâu, bản thân việc chuyển thế giới thành dữ liệu cũng kéo theo độ nghiêng. Vì vậy, điều chúng ta có thể làm không phải là xóa thiên lệch, mà là ngăn nó lan thành thiệt hại. SIPRI nói phải chặn ở hai nơi. Ở giai đoạn chế tạo, cần thu thập dữ liệu cân bằng và kiểm tra thiên lệch. Ở giai đoạn sử dụng, cần đưa sự kiểm soát và phán đoán của con người vào quy trình xác nhận mục tiêu.

Nhưng chính tại đây, chúng ta lại bị kéo trở về phần đầu tiên. Nếu người xác nhận mục tiêu chỉ có 20 giây, và nếu bên trong cỗ máy mà người ấy muốn nhìn vào là một hộp đen, thì sự kiểm soát của người ấy đang kiểm soát cái gì? Khuyến nghị đưa kiểm soát của con người vào là đúng. Nhưng sự kiểm soát của một người đã bị lấy mất thời gian và bị lấy mất cửa nhìn vào bên trong chỉ còn là kiểm soát trên danh nghĩa, rỗng ruột. Thiên lệch tự động hóa ru ngủ con người, hộp đen che mắt con người, áp lực thời gian đẩy tay con người đi. Khi cả ba cùng vận hành, sự kiểm soát của con người co lại thành một bàn tay đóng dấu.

Sức nặng thật sự của thế lưỡng nan hộp đen lộ ra sau sự việc. Sau khi chuyện ở Minab xảy ra, người ta hỏi: vì sao chuyện này xảy ra? Với một tai nạn thông thường, chúng ta có thể lần ngược nguyên nhân. Có thể chỉ ra ai đã làm sai điều gì, ở đâu. Nhưng tai nạn có hệ thống học sâu xen vào thì khác. Vì sao máy đưa tọa độ ấy lên thành mục tiêu, lý do ấy rải trong hàng triệu kết nối bên trong cỗ máy, nên con người khó kéo nó ra thành một câu nói. Người thiết kế cũng không dự đoán được, người vận hành cũng không giải thích được. Cái chết đã xảy ra, nhưng không có câu nào để giải thích cái chết ấy.

Một cái chết không thể giải thích. Cần nhai kỹ câu này một lần. Điều đầu tiên chúng ta muốn nghe trước cái chết của một ai đó là 'vì sao'. Cha mẹ mất con ở Minab cũng muốn hỏi điều ấy. Vì sao con tôi chết? Muốn trả lời câu hỏi đó, ai đó phải giải thích vì sao tọa độ ấy trở thành mục tiêu. Nhưng không có ai có thể đưa ra lời giải thích ấy. Máy không biết nói, con người không nhìn được vào bên trong máy. Câu hỏi 'vì sao' của cha mẹ treo lơ lửng trong khoảng không.

Đây là lý do nghịch lý hộp đen không dừng lại ở một khiếm khuyết kỹ thuật nhỏ. Nó là vấn đề của giải thích, và khi lời giải thích bị cắt đứt, trách nhiệm cũng bị cắt đứt theo. Nếu không thể chỉ ra nguyên nhân của sai lầm, ta cũng không thể xác định trách nhiệm về sai lầm ấy phải đặt lên ai. Thông tin rằng tọa độ của Minab đã là dữ liệu từ 10 năm trước cho ta biết một điều: ở đâu đó, dữ liệu đã không được cập nhật. Nhưng khoảnh khắc ta cố chỉ ra 'đâu đó' ấy là nơi nào, bàn tay của ai, ở bước nào trong hệ thống, ta lại đứng trước một chiếc hộp không nhìn thấy bên trong.

Cần nói rõ một điểm. Hộp đen không biến mọi cái chết thành điều không thể giải thích. Việc tọa độ của Minab đã là dữ liệu từ 10 năm trước ít nhất cho biết rằng ở một bước nào đó, dữ liệu đã cũ. Đến đó thì vẫn còn giải thích được. Nhưng ô tiếp theo lại trống. Ai lẽ ra phải cập nhật dữ liệu ấy, vì sao người đó không làm, vì sao hệ thống không lọc ra tọa độ đã cũ, khi lần theo từng mắt xích như vậy, con đường biến mất vào bên trong cỗ máy ở một điểm nào đó. Không phải hoàn toàn không có lời giải thích, mà lời giải thích bị cắt đứt trước khi chạm tới nơi có thể quy trách nhiệm. Sự đứt đoạn lưng chừng này còn khó chịu hơn. Nếu mọi thứ tối đen hoàn toàn, có lẽ ta còn có thể hỏi trách nhiệm của tất cả. Nhưng trong bóng tối chỉ hiện ra một nửa, mỗi người sẽ chỉ vào phần nhìn thấy trước mặt mình và nói phần còn lại là chuyện ở nơi không nhìn thấy.

Cái chết không thể giải thích dẫn tới cái chết không thể chịu trách nhiệm. Và nơi những cái chết không thể chịu trách nhiệm chất chồng lên nhau chính là khoảng trống mà cuốn sách này đã nói từ đầu. Phần tiếp theo sẽ nhìn vào hình dạng chính xác của khoảng trống ấy.

## 3. Khoảng trống trách nhiệm: không ai bóp cò

Hãy đặt thẳng câu hỏi. Trường học của Minab bị ném bom. Những đứa trẻ bảy tuổi đã chết. Vậy ai là người phải chịu trách nhiệm về việc này?

Hãy lần lượt đặt ba người lên trước mặt.

Trước hết là người tạo ra hệ thống. Đó là lập trình viên và nhà sản xuất. Họ đã tạo ra cỗ máy đề xuất tọa độ ấy làm mục tiêu. Nhưng họ có thể nói thế này. Chúng tôi không tạo ra nó để đánh vào trường học. Cỗ máy của chúng tôi chỉ được thiết kế để nhìn vào dữ liệu và tìm mẫu hình, còn việc dữ liệu đã cũ không phải trách nhiệm của chúng tôi. Hơn nữa, vì hệ thống deep learning tự học, ngay cả chúng tôi cũng không thể dự đoán chính xác nó sẽ chọn gì làm mục tiêu. Nếu người tạo ra nó không thể dự đoán kết quả, liệu có thể bắt người tạo ra chịu trách nhiệm về kết quả ấy không?

Tiếp theo là người ra lệnh cho chiến dịch. Đó là chỉ huy. Ông ta đã phê chuẩn chiến dịch ấy. Nhưng ông ta cũng có thể nói thế này. Tôi không ra lệnh đánh vào trường học. Tôi phê chuẩn mục tiêu do hệ thống đưa lên, và tôi không thể biết mục tiêu ấy thực ra là một trường học. Bên trong hệ thống cũng là hộp đen đối với tôi, tôi không thể dự liệu rằng tọa độ đã cũ, và tôi không thể dừng quả bom khi nó đã rơi.

Cuối cùng là chính cỗ máy. Nhưng không thể đưa cỗ máy ra trước tòa. Cỗ máy không có ý định hình sự. Nó không có ý muốn giết người, không có cảm giác tội lỗi, cũng không có hối hận. Chúng ta không thể trừng phạt cỗ máy, và với thứ không thể bị trừng phạt, chúng ta không thể đặt nó làm chủ thể chịu trách nhiệm.

Ba người đều đã được đưa ra, nhưng cuối cùng không còn ai cả. Người tạo ra nói rằng mình không thể dự đoán. Người ra lệnh nói rằng mình không thể thấy trước và cũng không thể dừng

lại. Cỗ máy thì ngay từ đầu đã không thể gánh lấy cái gọi là trách nhiệm. Thiệt hại bất hợp pháp đối với dân thường đã xảy ra, nhưng chỗ đứng để chịu trách nhiệm về thiệt hại ấy lại bỏ trống. Cái tên mà các học giả đặt cho chỗ trống này là responsibility gap, tức khoảng trống trách nhiệm.

Trước hết cần gỡ bỏ một hiểu lầm. Khi nghe nói đến khoảng trống trách nhiệm, người ta thường nghĩ đến chuyện của một tương lai xa. Đó là loại robot sát thương hoàn toàn tự chủ trong phim ảnh, tự chọn mục tiêu và tự bóp cò mà không cần mệnh lệnh của con người. Nếu là loại máy như vậy, việc xuất hiện một chỗ trống trong trách nhiệm có vẻ là điều đương nhiên. Nhưng chỗ trống mà cuốn sách này đã bàn đến không nằm xa như thế. Ở Gaza cũng vậy, ở Minab cũng vậy, trước cò súng vẫn có con người ngồi đó. Vũ khí hoàn toàn tự chủ không có mặt ở đó. Vậy mà chỗ đứng của trách nhiệm vẫn bỏ trống. Đây là điểm cốt lõi. Chỗ trống không xuất hiện vì con người biến mất. Nó xuất hiện khi con người vẫn ngồi ở đó, nhưng phán đoán của họ nhường chỗ cho phán đoán của máy. Có con người, nhưng không có trách nhiệm. Đó là nơi chúng ta đã đến rồi. Không phải lời cảnh báo về tương lai xa, mà là thực tế của năm 2026.

Nếu lần ngược về gốc rễ của khái niệm này, ta gặp cái tên Andreas Matthias. Ông là người đã chỉ ra rằng với kết quả do một cỗ máy tự học tạo ra, sẽ xuất hiện một khe hở mà cả người tạo ra lẫn cỗ máy đều không thể chịu trách nhiệm. Người đưa khe hở ấy vào vấn đề vũ khí tự chủ trong chiến tranh và dựng nó lên thật rõ là Robert Sparrow, triết gia của Đại học Monash ở Australia. Năm 2007, ông viết bài luận có tên 'Killer Robots'. Đó là bài viết truy đến cùng câu hỏi: khi vũ khí tự chủ gây ra hành vi tương đương tội ác chiến tranh, ai sẽ bị đưa ra trước tòa.

Lập luận của Sparrow không thừa lời, nhưng cũng không để lại lối thoát. Ông lần lượt xem xét ba ứng viên. Nếu đưa lập trình viên ra chịu trách nhiệm, thì máy càng tự chủ, hành vi của nó càng rời khỏi bàn tay người tạo ra. Sparrow viết như sau. "Khả năng một hệ thống tự chủ đưa ra lựa chọn mà lập trình viên của nó không dự đoán hoặc không khuyến nghị đã nằm sẵn trong chính tuyên bố rằng nó là tự chủ." Chính từ tự chủ đã có nghĩa là cắt sợi dây nối giữa người tạo ra và kết quả. Ông đưa ra một so sánh. Quy trách nhiệm cho lập trình viên về hành vi của một cỗ máy tự chủ cũng giống như quy trách nhiệm cho cha mẹ về hành vi của đứa con đã trưởng thành và rời khỏi vòng tay gia đình.

Vậy còn chỉ huy thì sao. Sparrow nói rằng đưa chỉ huy ra chịu trách nhiệm là câu trả lời mà quân đội thường nghiêng về. Hãy nghĩ đến pháo tầm xa. Dù quả đạn bay lệch và giết nhầm người, chỉ huy ra lệnh bắn khẩu pháo ấy vẫn chịu trách nhiệm. Ngay khoảnh khắc quyết định khai hỏa, ông ta gánh luôn cả nguy cơ bắn lệch. Nhưng Sparrow không dừng ở đây. Ông nói vũ khí tự chủ khác với khẩu pháo bắn lệch. Khẩu pháo bay về phía nơi ta đã nhắm, rồi do sai sót mà lệch đi. Vũ khí tự chủ tự chọn mục tiêu. Nếu cỗ máy thật sự tự chọn mục tiêu của mình, thì việc đổ toàn bộ cái chết phát sinh từ lựa chọn ấy lên chỉ huy là bất công. Bởi như vậy là trừng phạt ông ta vì một quyết định mà ông ta không kiểm soát được.

Nơi Sparrow đi đến là một ngõ cụt. Máy càng trở nên tự chủ, người tạo ra và người ra lệnh càng xa khỏi trách nhiệm, nhưng cũng không thể trừng phạt cỗ máy. Ông ví ngõ cụt này với lính trẻ em. Vì sao việc đưa trẻ em ra chiến trường lại khủng khiếp. Vì quyết định phân định sống chết được đặt vào tay một tồn tại không thể gánh trọn sức nặng của quyết định ấy. Khi một đơn vị trẻ em bước vào chiến trường, con người chết trong tình trạng không ai kiểm soát. Dù dân thường chết, cũng không có ai chịu trách nhiệm về cái chết ấy. Cái chết không phải ngẫu nhiên, nhưng cũng xảy ra như thể không thuộc trách nhiệm của bất kỳ ai. Sparrow cho rằng vũ khí tự chủ sẽ chiếm đúng vị trí đó.

Sparrow lấy từ nguyên tắc lâu đời của chiến tranh chính nghĩa để giải thích vì sao chỗ trống này lại nặng nề đến vậy. Đó là ý nghĩ rằng ngay cả trong chiến tranh cũng có những chuẩn mực phải giữ, tiếng Latin gọi là jus in bello. Sparrow nói rằng dưới đáy của nguyên tắc này có một điều kiện. Nếu đã giết người trong chiến tranh, phải có ai đó có thể chịu trách nhiệm về cái chết ấy. Ông gọi điều này là sự tôn trọng tối thiểu đối với kẻ thù. Câu văn của ông như sau. "Nếu chiến tranh chịu sự chi phối của đạo đức, thì việc ai đó chịu trách nhiệm, hoặc có thể chịu trách nhiệm, về quyết định lấy đi mạng sống của kẻ thù là mức tôn trọng tối thiểu mà kẻ thù đáng được nhận. Nếu không giữ được điều này, chúng ta đang đối xử với kẻ thù như sâu bọ. Như thể họ có thể bị tiêu diệt mà không cần bất kỳ cân nhắc đạo đức nào." Gia đình của người đã chết có tư cách được nghe câu trả lời. Vì sao người ấy chết, ai chịu trách nhiệm, lý do là gì. Không có người chịu trách nhiệm nghĩa là ngay từ đầu câu trả lời ấy đã không tồn tại.

Ở đây, có người có thể sẽ đưa ra một lối thoát. Chỉ cần giữ con người trong chuỗi quyết định là được, phải không. Dù máy móc chọn mục tiêu, nếu phát súng cuối cùng vẫn do con người bóp cò, khoảng trống trách nhiệm chẳng phải sẽ được lấp lại sao. Sparrow cho rằng con đường ấy cũng là ngõ cụt. Ông chỉ ra hai điều. Một là, khi tốc độ chiến đấu ngày càng nhanh, sẽ đến lúc thời gian con người cần để ra quyết định dài hơn thời gian được phép để sống sót. Khi ấy, đặt con người trong chuỗi quyết định không làm tăng an toàn, mà trái lại trở thành bất lợi. Hai là, dù con người có tiếng nói cuối cùng, nếu người đó dựa vào thông tin do máy đưa ra để quyết định, câu hỏi vẫn còn nguyên. Khi thông tin ấy sai, liệu có thể quy trọn trách nhiệm cho người đã ra quyết định đó không. Câu văn của Sparrow là thế này. "Câu hỏi ai chịu trách nhiệm đối với quyết định khác mà quyết định của người đó dựa trên, tức quyết định do trí tuệ nhân tạo đưa ra, vẫn nằm ở trung tâm." Chỉ đặt con người trong chuỗi quyết định không thể lấp được trách nhiệm đối với phán đoán của cỗ máy nằm ở phía trên chuỗi ấy. Điều chúng ta thấy ở Gaza và Iran chính là chuyện này. Con người có mặt trong chuỗi quyết định, nhưng điểm số và tọa độ mà người đó dựa vào lại do máy tạo ra.

Trong một bài viết khác năm 2016, "Robots and Respect", Sparrow đẩy ý nghĩ này đi thêm một bước. Trước hết, ông thẳng thắn chỉ ra rằng việc dùng vũ khí tự hành trong phạm vi hẹp có thể có nhiều dư địa được biện minh về mặt đạo đức hơn mức các nhà phê bình thừa nhận. Đó là thái độ không chỉ chọn những sự thật có lợi cho lập trường của mình. Ông đặt vũ khí tự hành ở

đâu đó giữa hai đầu mút. Một đầu là mìn sát thương chống người, nổ khi có áp lực đè lên. Đó là một thiết bị thô đến mức gọi là "quyết định" cũng thấy ngượng. Đầu kia là con người, hoặc một trí tuệ nhân tạo mạnh mà chúng ta chưa biết cách tạo ra. Điều ông đặt vấn đề nằm ở khoảng giữa. Đó là cỗ máy đứng ở vị trí lửng lơ, quá phức tạp để gọi là một quả mìn, nhưng quá thiếu hụt để gọi là con người. Định nghĩa ông đưa ra là thế này. Một hệ thống được giao nhiệm vụ nhận diện mục tiêu và chọn mục tiêu nào sẽ bị đánh, và ngay cả khi vận hành hoàn hảo vẫn còn một mức bất định nào đó về việc sẽ đánh gì và vì sao đánh. Lavender, Gospel, hệ thống đã hiển thị tọa độ của Minab, chính là đứng ở vị trí này.

Ngay cả sau đó, Sparrow vẫn giữ đến cùng trực giác rằng có điều gì đó sai. Nơi ông tìm thấy cái sai là sự tôn trọng. Giết kẻ địch bằng vũ khí tự hành là không thể hiện sự tôn trọng xứng đáng đối với nhân tính của kẻ địch. Một khi chúng ta chấp nhận rằng máy móc là thứ giết người, ở đó không có khuôn mặt của một con người khác đứng ra chịu trách nhiệm trước cái chết của một người. Ông thừa nhận rằng gốc rễ của trực giác này một phần dựa vào tập quán. Dù vậy, ông cho rằng điều chúng ta biểu đạt qua cách đối xử với kẻ địch có thể mang tính quyết định đối với đạo đức của chính cách đối xử ấy. Vì thế, ông kết luận rằng nền tảng cho lập luận vũ khí tự hành là vũ khí "xấu tự thân" không vững chắc như lời các nhà phê bình nói, nhưng vẫn đủ vững để cấm phát triển và triển khai loại vũ khí đó.

Khoảng trống mà Sparrow dựng lên bằng triết học đã được các luật gia chuyển sang ngôn ngữ cụ thể hơn. Năm 2015, Human Rights Watch và International Human Rights Clinic của Harvard Law School cùng công bố một báo cáo. Đó là bài viết "Mind the Gap" do Bonnie Docherty dẫn dắt. Báo cáo này lần theo khoảng trống trách nhiệm từng bước bằng ngôn ngữ pháp luật. Lập trình viên, nhà sản xuất, quân nhân đều có thể tránh được trách nhiệm. Nhìn vào trường hợp chỉ huy, trong tình huống thông thường, tức khi người đó không thể dự đoán thiệt hại và cũng không thể ngăn nó lại, người đó thoát khỏi trách nhiệm chỉ huy. Không thể dự đoán và không thể ngăn lại. Hai câu này là trọng tâm. Đó chính là lời mà chỉ huy của Minab đã nói.

Một câu trong báo cáo này tóm tắt ngắn nhất hệ quả của khoảng trống ấy. Không có trách nhiệm thì không có răn đe, cũng không có sự đền đáp đối với nạn nhân. Nghĩa là gì. Răn đe là tâm thế khiến người ta không dám làm bừa vì biết có ai đó sẽ bị trừng phạt. Nếu không có người chịu trách nhiệm, tâm thế ấy cũng không xuất hiện. Điều đó có nghĩa là lần sau cùng một việc vẫn có thể xảy ra. Đền đáp là sự đáp trả xứng đáng dành cho người bị hại. Nếu không có người chịu trách nhiệm, cha mẹ của Minab cũng không biết phải hướng đến đâu. Nỗi đau của họ không chạm được tới bất kỳ tòa án nào.

Giờ đây, chúng ta có thể thấy rõ hơn vì sao Sparrow ví khoảng trống này với trẻ em binh sĩ. Ông viết về trẻ em binh sĩ như sau. Các em đủ tự chủ để người lớn đưa các em ra chiến trường không thể kiểm soát hoàn toàn các em. Đồng thời, các em chưa đủ trưởng thành để tự chịu trách nhiệm về hành động của mình. Vì thế, khi một đơn vị trẻ em giết người, cái chết ấy lơ lửng

giữa không trung, không thuộc trách nhiệm của người lớn kiểm soát, cũng không thuộc trách nhiệm của đứa trẻ đã bóp cò. Sparrow gọi vị trí lửng lơ này là "tác nhân thông minh không có trách nhiệm đạo đức". Một tồn tại hành động thông minh nhưng không thể chịu trách nhiệm. Ông cho rằng vũ khí tự hành sẽ chiếm vị trí này trong một thời gian dài sắp tới. Hệ thống hiển thị tọa độ của Minab tinh vi đến mức khó chê. Nhưng sự tinh vi ấy là chuyện khác với trách nhiệm. Máy móc tinh vi cũng không thể chịu trách nhiệm.

Giờ có thể đặt rõ mệnh đề cốt lõi của cuốn sách này. Khi trí tuệ nhân tạo quân sự tự hành gây ra thiệt hại dân sự bất hợp pháp, chúng ta không thể hỏi máy móc về ý định hình sự. Máy móc không có ý định. Nhưng cũng khó đặt toàn bộ trách nhiệm ấy lên con người. Người tạo ra nó không thể dự đoán, người ra lệnh không thể lường trước và không thể ngăn lại. Trách nhiệm chảy xuống khe hở giữa con người và máy móc, rồi không đọng lại ở đâu cả.

Hãy quay lại chiếc bàn ở Gaza và màn hình ở Washington. Hai mươi giây phê chuẩn. 60 giây tấn công. Một nghìn tọa độ. Có một câu hỏi xuyên qua tất cả những cảnh ấy. Có thể nói rằng con người đã ra quyết định không. Giờ đây, chúng ta có thể thêm một lớp nữa vào câu hỏi đó. Khi quyết định ấy khiến một người bị giết oan, ai chịu trách nhiệm cho cái chết ấy.

Ở Gaza cũng vậy, ở Minab cũng vậy, thứ thả bom là máy bay do con người điều khiển, và ngón tay bấm nút phê chuẩn cũng là ngón tay của con người. Trước cò súng rõ ràng có người ngồi đó. Nhưng người ấy nhìn điểm số do máy đưa ra, nhìn tọa độ do máy chọn, rồi đóng dấu trong vòng hai mươi giây. Chúng ta không còn có thể phân biệt rành rọt rằng thứ người ấy kéo là cò súng, hay thứ người ấy bấm chỉ là con dấu đóng lên kết luận của máy.

Năm 2007, Sparrow đã nhìn thấy trước ngày việc này xảy ra và đi đến một kết luận. Nếu một vũ khí tạo ra cái chết mà không ai có thể chịu trách nhiệm, thì việc đưa vũ khí ấy vào chiến tranh là trái với điều kiện của một cuộc chiến chính nghĩa, nên không thể coi là có đạo đức. Ông cho rằng vũ khí tự chủ sẽ trở thành loại vũ khí như vậy. Khi ông viết bài vào năm 2007, đó còn là một giả định về tương lai. Khi ấy, ngay sau khi Lục quân Hoa Kỳ tuyên bố sẽ xây dựng một đội quân robot trước năm 2012, điều ông hình dung là những máy bay tự chọn mục tiêu. Hai mươi năm đã trôi qua, một phần giả định ấy đã thành hiện thực. Chỉ có hình dạng của nó hơi khác với điều ông vẽ ra. Nó không đến dưới dạng robot tự kéo cò, mà dưới dạng hệ thống đặt điểm số và tọa độ trước mặt con người. Và khoảng trống trách nhiệm đã đến trước, không chờ thứ vũ khí hoàn toàn tự chủ mà Sparrow từng giả định.

Nói rằng không ai kéo cò súng, thật ra là một cách nói khác của việc không ai chịu trách nhiệm. Chiếc ghế trước cò súng không hề trống. Có người ngồi ở đó. Chỉ là phán đoán của người ấy đã trống rỗng. Thứ trống không phải là chiếc ghế, mà là trách nhiệm.

Công nghệ không thể lấp khoảng trống ấy. Máy móc nhanh hơn cũng không lấp được, tọa độ chính xác hơn cũng không lấp được. Đó là việc pháp luật phải đảm nhận. Và pháp luật ấy hiện đã đi đến đâu, đang bỏ sót điều gì, chúng ta sẽ xem ở chương tiếp theo.

# Phần III Khoảng trống Chương 6 Luật nhân đạo quốc tế có thể buộc máy móc dừng lại không

## 1. Bối cảnh đã biến mất: cảm biến không đọc được cử chỉ đầu hàng

Một người lính đặt vũ khí xuống. Anh giơ hai tay lên quá đầu. Điều anh muốn nói rất rõ. Tôi sẽ không chiến đấu nữa, đừng bắn tôi. Nếu phía đối diện là một con người, người đó sẽ đọc được cử chỉ này trong 0,5 giây. Hai bàn tay đưa lên, khẩu súng rơi xuống, đôi vai khom lại. Người đó tiếp nhận tất cả cùng lúc và phán đoán rằng người này đang đầu hàng. Luật chiến tranh bảo vệ người lính ấy từ khoảnh khắc đó. Bởi vì việc bắn người đã bỏ vũ khí và đầu hàng bị cấm.

Bây giờ hãy nhìn lại cùng cảnh ấy bằng con mắt của máy móc. Thứ cảm biến nhìn thấy không phải là cử chỉ, mà là tín hiệu. Nhiệt, ánh sáng, chuyển động, hình dạng, tốc độ. Cảm biến bắt lấy đường nét của con người, bắt lấy động tác cánh tay đưa lên, rồi đem động tác ấy so với hồ sơ mục tiêu đã được nhập sẵn. Nhưng vấn đề nảy sinh ở đây. Động tác giơ hai tay để đầu hàng và động tác nâng súng lên vai để ngắm bắn, khi được quy đổi thành tín hiệu, lại giống nhau. Cả hai đều có cánh tay đưa lên. Cả hai đều có phần thân trên chuyển động. Hai cảnh hoàn toàn trái ngược đối với con người lại đi vào cảm biến như những dữ liệu tương tự nhau.

Nếu phía đối diện là một con người, trong 0,5 giây ấy người đó đọc được nhiều thứ hơn cùng một lúc. Ánh mắt của người lính, đôi môi run rẩy, sự do dự khi ném vũ khí, lời anh ta hét lên. Tất cả các tín hiệu ấy hợp lại thành một nghĩa gọi là 'đầu hàng'. Con người nhận ra nghĩa ấy với tốc độ gần như bản năng. Ngay cả một tân binh chưa từng tiếp nhận đầu hàng lần nào, khi thấy đối phương giơ hai tay và bỏ vũ khí, cũng rút ngón tay khỏi cò súng. Động tác rút ngón tay ấy chính là luật. Không cần ai dạy, con người không bắn người đã đầu hàng.

Đối với cảm biến, nghĩa ấy không tồn tại. Đối với cảm biến, chỉ có dữ liệu. Góc của cánh tay, nhiệt độ cơ thể, tốc độ chuyển động. Nó đem các con số này so với hồ sơ mục tiêu đã lưu sẵn, nếu khớp thì đánh dấu là có thể tấn công, nếu không khớp thì bỏ qua. Nghĩa của việc đầu hàng không có chỗ đứng ở bất kỳ đâu trong quá trình này. Nghĩa sinh ra từ bối cảnh, nhưng bối cảnh thoát mất ngay khi bị quy về con số.

Ví dụ này không phải do tôi nghĩ ra. Nó được ghi nguyên như vậy trong văn kiện lập trường mà Ủy ban Chữ thập đỏ Quốc tế công bố vào tháng 10 năm 2025. ICRC viết như sau. Hệ thống vũ khí tự chủ "có thể không phân biệt được giữa một người lính đặt vũ khí xuống và giơ hai tay để đầu hàng với một người lính nâng vũ khí lên vai để bắt đầu tấn công. Khi các động tác này được quy đổi thành đầu vào cảm biến, chúng có thể trông giống nhau đối với hệ thống, nhưng hệ quả pháp lý thì hoàn toàn khác." Cùng văn kiện ấy còn đi thêm một bước. Nó nói rằng cách biểu hiện ý nghĩa đầu hàng không chỉ có một. Có nước coi cờ trắng là đầu hàng, có nước chỉ coi đó là tín hiệu muốn thương lượng. Canada, Cote d'Ivoire và Hoa Kỳ coi cờ trắng là ý muốn thương lượng, còn Bỉ, Cameroon, Pháp và Cộng hòa Dominica coi đó là ý nghĩa đầu hàng. Cùng

một mảnh vải lại mang nghĩa khác nhau ở mỗi nước. Muốn đọc được khác biệt này cần có bối cảnh. Bối cảnh là việc của con người.

Nhan đề một bài luận mà ICRC trích dẫn nén lại sự căng thẳng của mục này. Đó là bài viết năm 2015 của triết gia Australia Robert Sparrow. Chính là Sparrow mà chúng ta đã gặp ở Chương 5. Nhan đề là 'Twenty Seconds to Comply'. Bài viết do Trường Chiến tranh Hải quân Hoa Kỳ xuất bản ấy đặt câu hỏi liệu vũ khí tự chủ có thể nhận ra một người lính đang đầu hàng hay không. Con số hai mươi giây trong nhan đề là câu hỏi liệu máy móc có thời gian để đáp lại người đang đầu hàng hay không. Ngay khi nhìn thấy con số này, chúng ta bị kéo về chiếc bàn ở Gaza. Ở đó, thời gian được trao để quyết định sự sống chết của một người cũng là hai mươi giây. Hai mươi giây ở một phía là thời gian cân nhắc có chấp nhận đầu hàng hay không, còn hai mươi giây ở phía kia là thời gian phê chuẩn mục tiêu. Cùng một con số rơi xuống hai đầu chiến trường với cùng một sức nặng.

Người lính bị thương cũng vậy. ICRC liệt kê những trạng thái mà một người lính bị thương có thể thể hiện. Anh ta có thể nằm bất tỉnh, có thể làm những cử động không thể hiểu được vì đau đớn không chịu nổi, cũng có thể bò về phía trận địa đối phương để được chữa trị. Trong tất cả các trường hợp ấy, người lính đó không còn là mục tiêu nữa. Luật chiến tranh bảo vệ anh ta. Nhưng một câu ICRC thêm vào đã chạm đúng trọng tâm của mục này. Tư cách của một con người có thể "chuyển từ mục tiêu hợp pháp sang người được bảo vệ chỉ trong vài giây." Chỉ vài giây. Liệu sự thay đổi xảy ra trong vài giây ấy có thể được đọc ra bằng các hồ sơ đã nhập sẵn hay không.

Luật nhân đạo quốc tế đứng trên ba nguyên tắc. Nguyên tắc thứ nhất là phân biệt. Phải phân biệt binh sĩ với dân thường, mục tiêu quân sự với cơ sở dân sự, và chỉ tấn công binh sĩ cùng mục tiêu quân sự. Nguyên tắc thứ hai là tương xứng. Thiệt hại dân sự phát sinh khi tấn công một mục tiêu quân sự không được vượt quá lợi ích quân sự thu được từ đòn tấn công đó. Nguyên tắc thứ ba là phòng ngừa khi tấn công. Trước khi tấn công, phải xác nhận mục tiêu có thật sự là mục tiêu quân sự hay không, phải thực hiện mọi biện pháp khả thi để giảm thiệt hại dân sự, và nếu phát hiện đó không phải là mục tiêu thì phải dừng tấn công. Cả ba nguyên tắc đều dựa trên cùng một tiền đề. Con người phải phán đoán theo từng tình huống cụ thể.

Đây chính là điều ICRC nhấn mạnh nhiều lần. Phán đoán này “không thể được giao cho xử lý bằng máy móc.” Việc xác định một vật có phải là mục tiêu quân sự hay không, một người có phải là dân thường được bảo vệ hay không, là phán đoán “không thể quy giản thành công thức toán học hay con số.” Giả sử một người lấy trộm vũ khí trong kho vũ khí. Nếu người đó lấy trộm vì lợi ích cá nhân, người đó vẫn là dân thường được bảo vệ. Nếu người đó lấy trộm để giúp địch, người đó trở thành mục tiêu trong thời gian hành vi ấy tiếp diễn. Cùng một động tác, nhưng ý định khác nhau. Phân định sự khác biệt ấy là việc của luật, và áp dụng luật ấy là việc của con người.

Bây giờ quay lại Minab. Đó là ngôi trường chúng ta đã thấy ở Chương 3, Trường tiểu học Shahr-e Tayebeh. Hôm ấy, tòa nhà đó được đánh dấu là cơ sở quân sự trên một bộ tọa độ cũ. Ta không biết nơi đó từng có ý nghĩa quân sự hay không. Nhưng khi tọa độ ấy hiện lên, tòa nhà đó là một trường tiểu học vừa kết thúc tiết một và bắt đầu tiết hai. Trẻ em đang ngồi trong lớp, các cô giáo đứng trước bảng. Nguyên tắc phân biệt hỏi: đây là mục tiêu quân sự hay cơ sở dân sự. Nguyên tắc tương xứng hỏi: lợi ích quân sự thu được ở đây có đủ để bù trừ những sinh mạng bên trong hay không. Nguyên tắc phòng ngừa hỏi: trước khi đánh, đã xác nhận tòa nhà ấy hiện vẫn là cơ sở quân sự hay chưa.

Để trả lời ba câu hỏi ấy, cần có bối cảnh. Tòa nhà đó hiện đang được dùng vào việc gì, bên trong có ai, tọa độ ấy được ghi từ khi nào. Mượn cách nói của ICRC, việc một vật có phải là mục tiêu quân sự hay không "phụ thuộc vào bối cảnh và bị ràng buộc bởi thời gian." Hôm qua là cơ sở quân sự không có nghĩa hôm nay vẫn là cơ sở quân sự. Tọa độ của Minab đã không theo kịp sự thay đổi ấy. Trang web của trường vẫn còn hoạt động, và ảnh vệ tinh có thể đã cho thấy sân trường. Nhưng hồ sơ được nhập vào hệ thống vẫn giữ nơi đó như một mục tiêu quân sự. Bối cảnh đã biến mất đâu đó. Ở chỗ bối cảnh biến mất ấy có 175 người.

Ví dụ mà ICRC đưa ra về đồ vật cũng chồng khít với Minab. Cùng văn bản lập trường đó viết như sau. Một xe dân sự có thể trở thành mục tiêu quân sự trong lúc tạm thời chở binh sĩ ra tiền tuyến, nhưng "sau khi việc sử dụng đó kết thúc, phá hủy nó không đem lại lợi ích quân sự nào." Tính chất quân sự của một vật không gắn cố định vào vật đó, mà xuất hiện rồi mất đi tùy theo hiện tại nó đang được dùng vào việc gì. Ngay cả nếu tòa nhà ở Minab từng có ý nghĩa quân sự, ý nghĩa ấy cũng không phải là một dấu hiệu vĩnh viễn. Khoảnh khắc nó trở lại làm trường học, tòa nhà ấy trở thành cơ sở dân sự. ICRC còn đi xa hơn và khẳng định rằng, nếu không đọc trước được những thay đổi như vậy, thì "không được phép phân loại trước một cách bao trùm các vật thể thành mục tiêu quân sự." Mục tiêu phải được xác nhận lại mỗi lần tấn công, chứ không phải đánh dấu một lần rồi dùng mãi. Tọa độ của Minab đã được dùng tiếp sau một lần đánh dấu như vậy.

Con số 10 phần trăm của Lavender cũng nằm đúng ở đó. Con số chúng ta đã thấy ở Chương 1, rằng khoảng 10 phần trăm những người máy đánh dấu là cần giết không phải là phần tử vũ trang. Những người bị đưa vào danh sách vì trùng tên, vì hồ sơ liên lạc giống nhau. Nguyên tắc phân biệt mà ICRC nói đến sụp xuống ngay ở con số 10 phần trăm ấy. Việc phân định một người là thành viên Hamas hay là nhân viên phòng vệ dân sự có cùng tên, đúng như ICRC nói, là phán đoán "không thể quy giản thành con số." Nhưng Lavender đã quy giản phán đoán đó thành điểm số từ 1 đến 100. Điểm số trông gọn gàng, và trong sự gọn gàng ấy, bối cảnh đã tuột ra ngoài.

Cần nói rõ một phân biệt ở đây. Lavender và Gospel không phải là hệ thống vũ khí tự động theo nghĩa hẹp mà ICRC định nghĩa. Báo cáo do Viện Nghiên cứu Hòa bình Quốc tế Stockholm,

Thụy Điển công bố vào tháng 6 năm 2025 tách hai loại này ra. Một bên là hệ thống vũ khí tự động. Đó là vũ khí sau khi được kích hoạt có thể tự chọn và tấn công mục tiêu mà không cần con người can thiệp thêm, tức hệ thống trong đó chính cò súng nằm bên trong máy. Bên kia là hệ thống trí tuệ nhân tạo hỗ trợ ra quyết định. Hệ thống này đề xuất mục tiêu và đưa ra điểm số, nhưng để cò súng lại cho con người. Lavender gần với loại thứ hai hơn. Vì con người đã đóng dấu phê chuẩn.

Giữa hai hệ thống có khác biệt trong cách vận hành, nhưng kết quả chúng tạo ra có điểm giống nhau. Điều ICRC cảnh báo nặng nề nhất về hệ thống vũ khí tự động là tính không thể dự đoán. Một hệ thống vận hành bằng học máy, theo cách nói của ICRC, ngay cả sau khi bắt đầu hoạt động, "chức năng của nó có thể thay đổi, và nó có thể dùng lực theo cách con người không lường trước trong những tình huống con người không thấy trước." Một hệ thống mà con người không hiểu, không dự đoán được, cũng không giải thích được tác động của nó. ICRC gọi hệ thống như vậy là "về bản chất mang tính không phân biệt." Không phân biệt có nghĩa là tấn công mà không phân biệt binh sĩ và dân thường. Vũ khí không thể đánh có phân biệt thì theo luật nhân đạo quốc tế hiện hành cũng đã bị cấm.

Lavender đã dừng hoạt động và chuyển điểm số cho con người, nên nó khác về bản chất với tính khó đoán mà Ủy ban Chữ thập đỏ nói đến. Nhưng chúng có chung một điều. Con người không thể giải thích đầu ra ấy. Vì sao Lavender chấm người đó 90 điểm, vì sao nó kéo vào một nhân viên dân phòng có cùng tên, sĩ quan tình báo đặt dấu phê chuẩn không thể biết được. Sự mờ đục mà Ủy ban Chữ thập đỏ gọi là hộp đen tụ lại ở cùng một chỗ, bất kể đó là hệ thống vũ khí tự trị hay hệ thống hỗ trợ ra quyết định. Con dấu đóng lên một điểm số không thể giải thích khiến khi điểm số ấy sai, người ta cũng không thể giải thích trách nhiệm thuộc về ai.

Nhưng sự phân biệt này không đem lại an ủi. Trái lại, nó đặt ra một câu hỏi khó hơn. Hệ thống vũ khí tự trị ít nhất đang đứng giữa tâm điểm tranh luận. Đó chính là đối tượng mà Ủy ban Chữ thập đỏ kêu gọi cấm, và Tổng thư ký Liên Hợp Quốc cũng kêu gọi cấm. Nhưng các hệ thống hỗ trợ ra quyết định như Lavender lại lách khỏi trung tâm của cuộc tranh luận cấm đoán ấy. Ngay cả tuyên bố lập trường năm 2025 của Ủy ban Chữ thập đỏ cũng ghi rõ rằng hệ thống hỗ trợ ra quyết định nằm "ngoài phạm vi của báo cáo này". Vì cò súng vẫn nằm trong tay con người, nên nó không phải vũ khí tự trị. Việc con người nằm trong chuỗi và việc con người thật sự phán đoán không phải là một. Chúng ta đã nói điều này ở Chương 3. Khi sĩ quan tình báo của Minav được cho 1 phút cho mỗi tọa độ, khi sĩ quan tình báo ở Gaza được cho 20 giây cho mỗi mục tiêu, người đó vẫn nằm trong chuỗi. Chỉ là phán đoán của anh ta đã rỗng. Điều luật muốn bảo vệ không phải vị trí của con người, mà là phán đoán của con người; nhưng tấm lưới của pháp luật bắt được vị trí và để lọt phán đoán.

## 2. Sự kiểm soát có ý nghĩa của con người

Người đầu tiên chỉ đúng khoảng trống này không phải chính phủ, cũng không phải quân đội. Đó là một tổ chức phi lợi nhuận nhỏ ở Anh. Tổ chức ấy tên là Article 36. Cái tên lấy từ Điều 36 của Nghị định thư bổ sung thứ nhất của Công ước Geneva. Điều khoản này quy định rằng khi phát triển hoặc đưa vào dùng một loại vũ khí mới, phải xem xét nó có phù hợp với luật quốc tế hay không. Khoảng năm 2013, tổ chức này đã tạo ra một cụm từ: "sự kiểm soát có ý nghĩa của con người đối với từng cuộc tấn công riêng lẻ" (meaningful human control over individual attacks).

Cụm từ này lan rất nhanh. Nhiều nước đã dùng nó tại các cuộc họp của Liên Hợp Quốc, và nó cũng xuất hiện trong thư ngỏ do các nhà nghiên cứu trí tuệ nhân tạo ký tên. Nhưng khi cụm từ ấy lan ra, phê phán cũng đi theo. Người ta nói chữ "có ý nghĩa" quá mơ hồ. Ai sẽ quyết định đâu là kiểm soát có ý nghĩa, đâu là kiểm soát không có ý nghĩa? Richard Moyes của Article 36 cho rằng phê phán này đã bỏ lỡ trọng tâm. Tính từ "có ý nghĩa" không phải kết luận, mà là điểm xuất phát. Nó là tín hiệu yêu cầu lấp đầy chỗ trống. Nó là lời mời để cộng đồng quốc tế cùng định nghĩa kiểm soát của con người phải có hình thức nào.

Ý nghĩ mà Moyes đặt dưới đáy khái niệm kiểm soát có ý nghĩa ngắn nhưng chắc. Nó bắt đầu từ hai tiền đề. Một, cỗ máy dùng vũ lực mà hoàn toàn không có sự kiểm soát của con người là điều không thể chấp nhận. Hai, việc con người nhìn dấu hiệu do máy tính đưa ra rồi chỉ bấm nút "phóng", bấm mà không có sự rõ ràng về nhận thức, không có ý thức, thì không thể gọi là kiểm soát có ý nghĩa. Hãy đọc lại tiền đề thứ hai này. Con người nhìn dấu hiệu của máy tính rồi chỉ bấm nút. 20 giây ở Gaza, 1 phút của Minav, đều nằm trong câu đó. Ngay từ năm 2013, Moyes đã vẽ sẵn cảnh ấy.

Vậy để kiểm soát có ý nghĩa, cần có những gì? Article 36 nêu ra bốn yếu tố cốt lõi. Một, công nghệ có thể dự đoán, đáng tin cậy và minh bạch. Người dùng phải hiểu được cỗ máy ấy hoạt động như thế nào. Hai, thông tin chính xác được trao cho người dùng. Đó là thông tin về mục tiêu muốn đạt được, về bản thân công nghệ ấy, và về bối cảnh nó được dùng. Ba, phán đoán và hành động kịp thời của con người, cùng khoảng trống để can thiệp đúng lúc. Bốn, trách nhiệm giải trình. Phải có cấu trúc để truy trách nhiệm khi sự việc sai đi.

Center for a New American Security của Hoa Kỳ chuyển cùng ý nghĩ ấy sang ngôn ngữ quân sự, rồi sắp xếp kiểm soát có ý nghĩa thành ba thành phần. Một, con người phải đưa ra quyết định có ý thức, dựa trên thông tin, về việc dùng vũ khí. Hai, con người phải có đủ thông tin để xác nhận tính hợp pháp của hành động của mình. Ba, vũ khí phải được thiết kế và thử nghiệm đúng cách, con người phải được huấn luyện đúng cách, để nó hoạt động như đã định. Hãy đọc lại hai từ trong yếu tố đầu tiên: dựa trên thông tin, có ý thức. Có ngón tay bấm nút không có nghĩa quyết định ấy có ý thức. Nếu tọa độ hiện lên và thời gian chỉ có 1 phút, ngón tay ấy vẫn bấm, nhưng ý thức không theo kịp.

Hãy đem bốn yếu tố này đặt lên Minav. Đó có phải là một công nghệ có thể dự đoán và minh bạch không? Con người có thể nhìn vào và hiểu vì sao hệ thống ấy đưa ngôi trường đó lên thành mục tiêu quân sự không? Thông tin chính xác có được cung cấp không? Sĩ quan tình báo có thể biết tọa độ ấy được ghi từ khi nào, và hiện nay tòa nhà đó đang được dùng vào việc gì không? Có thể phán đoán và can thiệp đúng lúc không? Với 1 phút cho mỗi tọa độ, với tốc độ xử lý 1.000 mục tiêu trong 24 giờ, con người có thể dừng lại và hỏi: "Khoan đã, tọa độ này có đúng không" không? Có thể quy trách nhiệm không? Ở Chương 5, chúng ta đã thấy câu trả lời. Người tạo ra nói rằng họ không thể dự đoán. Người ra lệnh nói rằng họ không thể lường trước và không thể dừng lại. Cả bốn yếu tố đều trống rỗng.

Article 36 còn chỉ ra thêm một điều. Kiểm soát có ý nghĩa không được quyết định chỉ bằng năng lực kỹ thuật. Cùng một cỗ máy, nếu dùng trong một khu vực hẹp và trong thời gian ngắn thì có thể kiểm soát được; nhưng nếu thả nó trong một khu vực rộng và trong thời gian dài, kiểm soát sẽ mờ đi. Vì trong thời gian ấy, con người và đồ vật ra vào khu vực đó. Một ngôi trường trống khi hệ thống bắt đầu vận hành có thể đầy trẻ em một giờ sau đó. Vì thế, diện tích càng rộng và thời gian càng dài, kiểm soát càng giảm. Nhận định này chạm thẳng vào con số 1.000 mục tiêu và mốc 24 giờ. Quy mô càng lớn, tốc độ càng nhanh, kiểm soát càng loãng.

Các học giả đã làm cho ý tưởng này sắc hơn. Filippo Santoni de Sio và Jeroen van den Hoven chia kiểm soát có ý nghĩa thành hai điều kiện: tracking và tracing. Tracking nghĩa là hệ thống phải phản ứng đúng với ý định và phán đoán của người kiểm soát. Khi con người bảo dừng, nó phải dừng; nó chỉ được vận hành trong những giới hạn do con người đặt ra. Tracing nghĩa là nếu lần ngược hành vi của hệ thống, ta phải đi tới được một con người có thể hiểu và chịu trách nhiệm về hành vi đó. Ở cuối hành vi phải có khuôn mặt con người. Nhưng khi lần ngược một đòn tấn công của Minav, thay vì khuôn mặt con người, ta chạm tới một tọa độ cũ. Nơi chuỗi truy ngược bị đứt, đó chính là khoảng trống.

Vì sao hai điều kiện này quan trọng, học giả luật người Nam Phi Thompson Chengeta đã nói gọn trong một câu. Trong bài viết đăng trên New York University Journal of International Law, ông hỏi lý do lớn nhất khiến ta bàn về kiểm soát có ý nghĩa là gì, rồi tự trả lời: bởi vũ khí tự chủ có thể "tạo ra một khoảng trống trách nhiệm pháp lý". Từ ông dùng là vacuum. Đó chính là thứ mà suốt cuốn sách này chúng ta gọi là chỗ trống. Cách xử lý của Chengeta rất rõ. Tính chất kiểm soát của con người phải bảo đảm rằng người đó có thể "chịu trách nhiệm về mọi hành động sau đó của robot". Mục đích của việc định nghĩa kiểm soát chính là giữ trách nhiệm lại. Khi kiểm soát mờ đi, trách nhiệm tuột xuống; nơi trách nhiệm tuột xuống là khoảng trống. Kiểm soát có ý nghĩa là một thiết kế nhằm chặn khoảng trống ấy từ trước.

Cùng cuốn nhập môn đó kéo cuộc bàn luận này thêm một bước xuống thực tế vận hành quân sự. Nó nói rằng kiểm soát không phải là việc xảy ra ở đầu ngón tay của một người. Người thiết kế vũ khí, người thử nghiệm, người đặt ra quy tắc tác chiến, người tạo thông tin mục tiêu, và

cuối cùng là người bấm nút, tất cả hợp thành một hệ thống kiểm soát. Nhưng câu này có hai mặt. Vì kiểm soát được chia cho nhiều người cũng có nghĩa là trách nhiệm được chia cho nhiều người. Đúng như chúng ta đã thấy ở Chương 5. Ai cũng cầm một mảnh của quyết định, nhưng không ai cầm toàn bộ quyết định. Vấn đề mà khái niệm kiểm soát có ý nghĩa muốn giải quyết chính là sự phân tán này. Nó muốn gom kiểm soát về lại một điểm, để bịt khoảng hở nơi trách nhiệm có thể tuột xuống.

Ở đây, tôi sẽ nói rõ lập trường. Lập luận cho rằng chỉ cần con người nằm trong chuỗi là đã đáp ứng kiểm soát có ý nghĩa là điều khó chấp nhận. Lập luận ấy nhầm vị trí với kiểm soát. Nó xem việc có một người ngồi trên ghế và việc người đó đang phán đoán là cùng một chuyện. Gaza và Minav cho thấy chính xác rằng hai chuyện ấy khác nhau. Thước đo kiểm soát có ý nghĩa là thước đo sự khác biệt đó. Nó không hỏi con người có ngồi đó không, mà hỏi con người có thời gian, thông tin và thẩm quyền để phán đoán không. Nếu đo bằng thước này, nhiều cảnh chúng ta đã thấy không đạt yêu cầu kiểm soát.

Ở đây ta phải đối diện với một sự thật khó chịu. Muốn giữ kiểm soát có ý nghĩa, phải có người phụ trách kiểm soát ấy. Trong thủ tục xác định mục tiêu, đó là người xem xét pháp luật, tức luật sư quân sự. Nhưng có các bản tin cho thấy vị trí của những người này đang lung lay. Bộ trưởng Quốc phòng Mỹ Pete Hegseth được cho là đã nêu ra khung đối lập 'sát thương tối đa đối đầu tính hợp pháp nhạt nhẽo' (maximum lethality vs tepid legality). Tinh thần ở đây là nhiệm vụ của quân đội là giết kẻ địch, chứ không để tư vấn pháp lý níu chân. Theo các bản tin của Just Security và Military.com, ông đã thay thế những người đứng đầu tổ chức pháp vụ quân đội, và đi theo hướng làm yếu vai trò giám sát của luật sư quân sự trong thủ tục xác định mục tiêu.

Trong đoạn này có một nghịch lý. Nhà địa lý học người Anh Craig Jones, trong cuốn 'The War Lawyers', đã vẽ ra hai khuôn mặt của luật sư quân sự. Một mặt, luật sư quân sự là người hợp pháp hóa bạo lực. Họ đóng dấu pháp luật lên thủ tục xác định mục tiêu, biến đòn tấn công ấy thành hợp pháp. Theo một nghĩa nào đó, họ khiến nhiều đòn tấn công hơn trở nên khả thi. Vì khi có sự bảo đảm về tính hợp pháp, quân đội có thể bóp cò tự tin hơn. Nhưng mặt khác, chính việc luật sư ấy có mặt trong thủ tục cũng là một thiết bị hãm. Việc từng có một vị trí để ông ta nói "mục tiêu này không được" đã đủ để một điều gì đó được lọc lại một lần.

Điểm nguy hiểm trong đường hướng của Hegseth nằm ở đây. Tốc độ xác định mục tiêu đang được máy móc đẩy lên, nhưng nếu ngay cả con người từng có thể buộc tốc độ ấy dừng lại một nhịp cũng bị đẩy ra khỏi thủ tục, khoảng trống sẽ rộng hơn. Máy móc đặt xuống tọa độ nhanh hơn, còn người chất vấn tọa độ ấy lại ít đi. Khe hở mà khái niệm kiểm soát có ý nghĩa muốn lấp lại đang bị nới rộng theo hướng ngược lại.

Ở đây còn có một cái bẫy tinh vi hơn. Khi luật sư quân sự đóng dấu vào quy trình lựa chọn mục tiêu và biến cuộc tấn công ấy thành hợp pháp, con dấu đó cũng có thể trở thành cái cớ để chuyển gánh nặng trách nhiệm đi nơi khác. Khi có chuyện sai, chỉ một câu 'đã qua thẩm tra

pháp lý' sẽ thành lá chắn. Vì con người đã xem xét và nói rằng hợp pháp, nên bề ngoài nó trở thành phán đoán của con người. Nhưng nếu người ấy chỉ lướt qua điểm số và tọa độ do máy đưa ra trong vòng một phút rồi đóng dấu, cái bề ngoài ấy rỗng ruột. Con dấu pháp lý trông như đã lấp đầy vị trí trách nhiệm, nhưng thật ra có thể chỉ là chiếc nắp mang tên hợp pháp đậy lên một khoảng trống. Điều kiểm soát có ý nghĩa hỏi không phải là đã có con dấu hay chưa, mà là dưới con dấu ấy có phán đoán thật hay không.

## 3. Điều khoản Martens và lương tri của nhân loại

Luật có những khoảng trống. Khi một loại vũ khí mới xuất hiện mà chưa có điều khoản nào trực tiếp cấm nó, vũ khí ấy lọt qua mắt lưới của luật. Không có trong điều khoản, nên không bị cấm; vì vậy có thể dùng. Để chặn lối lập luận này, luật nhân đạo quốc tế từ lâu đã đặt sẵn một thiết bị an toàn. Đó là Điều khoản Martens (Martens Clause).

Nơi điều khoản này ra đời có một lịch sử nhỏ. Tại Hội nghị Hòa bình La Hay năm 1899, các cường quốc và các nước yếu hơn đã va chạm quanh một vấn đề. Đó là phải đối xử thế nào với cư dân cầm vũ khí trong vùng bị chiếm đóng. Khi thỏa thuận bị bế tắc, luật gia Fyodor Martens, đại diện của Nga, đưa ra một câu. Đó là câu trả lời cho việc phải lấp khoảng trống không có quy định thành văn bằng gì. Thứ ông đặt vào đó không phải một điều khoản mới, mà là một chuẩn mực nằm ngoài văn bản điều luật.

Cốt lõi của câu chữ ông trau chuốt là như sau. Ngay cả trong những trường hợp quy định thành văn không xử lý, dân thường và chiến binh vẫn nằm dưới sự bảo vệ của các nguyên tắc luật quốc tế phát sinh từ 'các tập quán đã được xác lập, từ các nguyên tắc nhân đạo, từ mệnh lệnh của lương tri công cộng'. Nói dễ hiểu là thế này. Không phải cứ không được ghi rõ trong bộ luật thì muốn làm gì cũng được. Chuẩn mực lương tri mà nhân loại đã tích lũy qua thời gian dài, các nguyên tắc nhân đạo, vẫn bảo vệ con người. Sự im lặng của điều luật không phải là tín hiệu cho phép.

Đây là lý do điều khoản này trở nên quan trọng trong tranh luận về vũ khí sát thương tự chủ. Chưa có điều ước nào trực tiếp cấm vũ khí tự chủ. Luật nhân đạo quốc tế không nhắc rõ đến nó dù chỉ một chữ. Vì vậy có người nói rằng vì không có điều khoản cấm, vũ khí tự chủ là hợp pháp. Điều khoản Martens chặn thẳng lời nói ấy. Dù không có điều khoản, các nguyên tắc nhân đạo và mệnh lệnh của lương tri công cộng vẫn đo lường loại vũ khí này. ICRC nhiều lần kéo điểm này trở lại. Vấn đề sâu nhất mà hệ thống vũ khí tự chủ đặt ra là vấn đề đạo đức trước khi là vấn đề pháp lý. Theo cách nói của ICRC, vũ khí tự chủ 'thay thế quyết định của con người về sự sống và cái chết bằng cảm biến, phần mềm và xử lý của máy móc'. Và ICRC nói rằng có một điều mà phần lớn con người đồng ý. Thuật toán, việc xử lý của máy móc, không được quyết định ai sống và ai chết.

ICRC chuyển trực giác đạo đức này thành các khuyến nghị cụ thể. Từ lập trường tháng 5 năm 2021 đến tuyên bố lập trường năm 2025, cốt lõi ấy không dao động. ICRC nói rằng 'cần khẩn cấp có các quy tắc mới có tính ràng buộc pháp lý'. Và họ chia đường nét của các quy tắc ấy thành ba phần. Thứ nhất, các hệ thống vũ khí tự chủ không thể dự đoán phải bị loại trừ rõ ràng. Chuyển nguyên văn câu của ICRC thì là: 'Các hệ thống vũ khí tự chủ không thể dự đoán phải bị loại trừ rõ ràng, nhất là vì tác động bừa bãi của chúng (Unpredictable autonomous weapon systems should be expressly ruled out, notably because of their indiscriminate effects).' Một vũ khí mà người sử dụng không thể hiểu đầy đủ, dự đoán và giải thích tác động của nó, theo lời ICRC, giống như 'bắn khi bị bịt mắt'. Hộp đen mà chúng ta đã thấy ở Chương 5 đứng đúng ở đây. Một hệ thống không thể giải thích vì sao người đó trở thành mục tiêu thì có tính bừa bãi, nên ngay cả theo luật hiện hành cũng đã bị cấm.

Thứ hai, vũ khí tự động trực tiếp nhắm vào con người, tức vũ khí tự động chống người, phải bị cấm. Lời ICRC gắn với khuyến nghị này có sức nặng. Việc lấy con người làm mục tiêu là "quy giản sinh mạng con người thành dữ liệu cảm biến và xử lý máy móc", và đó là "cái chết bằng thuật toán, bước chuyển cuối cùng trong tự động hóa sát thương". Điều chúng ta đã thấy xuyên suốt cuốn sách này chính là sự quy giản ấy. Khi Lavender quy giản một con người thành điểm số từ 1 đến 100, khi tọa độ của Minrav quy giản 175 người thành một dòng dữ liệu, ở cuối sự quy giản đó là đường ranh mà ICRC đã vạch ra.

Thứ ba, các vũ khí tự động còn lại phải bị hạn chế nghiêm ngặt. Nghĩa là phải thu hẹp loại mục tiêu vào các mục tiêu quân sự đúng nghĩa, giới hạn thời gian, phạm vi địa lý và quy mô của chiến dịch, chỉ cho dùng trong tình huống không có dân thường, và bảo đảm con người có thể giám sát theo thời gian thực, can thiệp kịp thời và dừng hoạt động của hệ thống. Đọc lại mục thứ ba này, ta thấy nó gần như chồng lên bốn yếu tố của kiểm soát có ý nghĩa đã nêu ở phần trước. Khuyến nghị pháp lý của ICRC và khung đạo đức của Article 36 xuất phát từ hai con đường khác nhau, rồi gặp nhau ở cùng một chỗ.

Phần ICRC chỉ ra nguyên tắc tương xứng và nguyên tắc phòng ngừa lại nhắm thẳng vào Minrav một lần nữa. Nguyên tắc tương xứng, theo cách nói của ICRC, đòi hỏi một phán đoán "trước khi xảy ra" (ex-ante). Trước khi đánh, tức trước khi sự việc xảy ra, phải cân nhắc thiệt hại dân sự dự kiến và so với lợi ích quân sự. Nhưng với vũ khí tự động, vào thời điểm kích hoạt, người ta không biết ai, khi nào, ở đâu sẽ trở thành mục tiêu. Vì vậy phán đoán trước khi xảy ra bị lệch. Nguyên tắc phòng ngừa còn trực tiếp hơn. ICRC viết như sau. Chỉ những chỉ thị được đưa ra trước cuộc tấn công "tự thân không đủ để trở thành biện pháp phòng ngừa đầy đủ". Bởi nếu lộ ra rằng mục tiêu không phải là mục tiêu quân sự, nghĩa vụ hủy bỏ hoặc dừng cuộc tấn công vẫn còn đến tận cuối cùng. ICRC nói rằng thông tin phải được thu thập và cập nhật "cho đến thời điểm mở màn cuộc tấn công, và trong chừng mực có thể, cả trong khi cuộc tấn công đang diễn ra". Tọa độ của Minrav là thông tin cũ không trải qua sự cập nhật ấy. Một chỉ thị đã được nhập một lần rơi xuống thành bom mà không có cơ hội gặp tín hiệu cho thấy nó đã sai.

Trên sân khấu quốc tế cũng có chuyển động theo cùng hướng. Tổng thư ký Liên Hợp Quốc Antonio Guterres đã để lại một câu trong báo cáo năm 2024. Bản thân báo cáo ấy có sức nặng. Vì đó là văn kiện chính thức do Tổng thư ký tập hợp ý kiến của các quốc gia thành viên và quốc gia quan sát viên theo yêu cầu của Đại hội đồng Liên Hợp Quốc. Trong đó, Guterres viết như sau. "Máy móc có quyền lực và quyền tùy nghi để tước đi sinh mạng con người là điều không thể chấp nhận về chính trị, ghê tởm về đạo đức, và phải bị cấm theo luật quốc tế." Ông còn đặt mốc thời gian. Đến năm 2026, cần ký kết một văn kiện có tính ràng buộc pháp lý, cấm các vũ khí tự động vận hành không có kiểm soát của con người và không thể tuân thủ luật nhân đạo quốc tế, đồng thời quản lý các loại còn lại. Cùng báo cáo ấy còn thêm một câu nữa. "Việc máy móc tự động lấy con người làm mục tiêu là một ranh giới đạo đức không được vượt qua."

Dòng chảy này đang có lực đẩy, và điều đó cũng hiện ra trong các con số. Theo tổng hợp của UN News và American Society of International Law, cuối năm 2023, tại Đại hội đồng Liên Hợp Quốc, 156 quốc gia đã bỏ phiếu thuận cho nghị quyết về vũ khí tự động. Đó là đa số áp đảo. Nhưng để công bằng, cũng phải ghi lại những tiếng nói thận trọng. Trong số các nước gửi ý kiến cho báo cáo của Tổng thư ký, một số cho rằng chưa cần đến một văn kiện ràng buộc mới, vì luật nhân đạo quốc tế hiện có đã đủ. Một số cường quốc quân sự coi vũ khí tự động là tài sản chiến lược và muốn làm chậm tốc độ đàm phán. Nhóm chuyên gia chính phủ theo Công ước về một số vũ khí thông thường (CCW GGE) đã thảo luận trong nhiều năm, nhưng trong khuôn khổ đòi hỏi đồng thuận tuyệt đối ấy, thỏa thuận vẫn tiến rất chậm. Lá phiếu thuận của 156 quốc gia và lực hãm của vài cường quốc cùng tồn tại trên một sân khấu.

Vậy Hàn Quốc đứng ở đâu trên sân khấu này. Báo cáo có tên 'Thời đại Next Oppenheimer' do Kim Hyun-jung thuộc Institute for National Security Strategy viết năm 2024 cho thấy góc nhìn của Hàn Quốc. Oppenheimer trong tiêu đề là nhà khoa học đã tạo ra bom hạt nhân. Sau khi chứng kiến Hiroshima và Nagasaki, ông trở thành một trong những người đi đầu trong kiểm soát vũ khí đối với thứ vũ khí do chính mình tạo ra. Báo cáo này cho rằng trong thời đại vũ khí sát thương tự động, một khoảnh khắc tương tự, tức 'Next Oppenheimer Moment', đang đến gần. Nhưng hướng mà báo cáo khuyến nghị lại khác giọng với lập luận cấm đoán của ICRC hay Guterres.

Lập luận của báo cáo như sau. Trong thời đại hạt nhân trước đây, thế giới bị chia thành những nước có hạt nhân và những nước không có hạt nhân. Các nước có hạt nhân trở thành bên tham gia đàm phán kiểm soát vũ khí, còn các nước không có hạt nhân chỉ còn là đối tượng thụ động của chế độ không phổ biến. Muốn ngồi vào bàn đàm phán thì trước hết phải có thứ vũ khí ấy. Nếu chuyển bài học này sang vũ khí tự động, kết luận là để Hàn Quốc đứng ở vị trí dẫn dắt trong thảo luận kiểm soát vũ khí, Hàn Quốc phải phát triển và đưa vào dùng vũ khí sát thương tự động trước. Đừng ở lại vị trí đối tượng của không phổ biến, mà hãy trở thành bên tham gia kiểm soát. Báo cáo nêu thực tế khó bố trí binh lực gần đường đình chiến vì dân số giảm mạnh, rồi đề xuất trước hết triển khai ở tiền tuyến các vũ khí tự động có chức năng giám

sát và trinh sát, sau đó sửa dần những điểm còn thiếu. Báo cáo còn dẫn cả câu châm ngôn 'First come, first served'.

Chỉ lên án góc nhìn này rồi bỏ qua thì không thật thà. Trong đó có sức nặng của thực tế. Quân số giảm, điều kiện của đường đình chiến, vị thế của một quốc gia tầm trung kẹt giữa các cường quốc. Phép tính lạnh rằng sức đàm phán cuối cùng đến từ thứ mình đang có cũng không sai. Nhưng nếu đặt cạnh những gì cuốn sách này đã thấy ở Gaza, ở Minrav, một câu hỏi hiện lên. Lập luận 'phát triển trước, sửa sau' ấy quá giống với một từ mà chúng ta đã theo suốt cuốn sách này. lethal beta. Nguyên mẫu chết người. Một thứ vũ khí chưa hoàn tất đã ra chiến trường, tích lũy dữ liệu trong thực chiến rồi được sửa tiếp. Tọa độ của Minrav bị cũ cũng vì hệ thống ấy chưa hoàn chỉnh. Nói triển khai trước rồi sửa những điểm còn thiếu cũng có nghĩa là trước khi những thiếu sót ấy được lấp đầy, sẽ có ai đó phải trả giá cho chúng. 175 người của Minrav đã trả cái giá ấy.

Một đoạn khác của báo cáo chia mức độ can thiệp của con người trong vận hành vũ khí tự động thành ba cấp. Can thiệp chủ động, khi con người phê chuẩn mọi bước; can thiệp thụ động, khi con người chỉ can thiệp vào các hành động quan trọng; và tự động hoàn toàn, khi con người bị loại ra. Báo cáo còn đưa ra một dự báo. Với vũ khí cấp chiến lược, vốn nhạy cảm với cán cân quyền lực giữa các quốc gia, sự can thiệp của con người sẽ được đòi hỏi; nhưng ngoài phạm vi đó, súng cá nhân hoặc vũ khí cấp đơn vị nhỏ có thể được phát triển theo hướng tự động hoàn toàn. Nghĩa là tay con người vẫn đặt trên vũ khí chiến lược, còn vũ khí nhỏ có thể được thả ra. Nếu dự báo này đúng, những loại vũ khí phổ biến nhất trực tiếp nhắm vào con người lại chính là những thứ đầu tiên thoát khỏi sự kiểm soát của con người. Chính thứ mà ICRC muốn cấm trước hết: vũ khí trực tiếp lấy con người làm mục tiêu.

Lựa chọn của Hàn Quốc không phải là điều cuốn sách này có thể quyết định thay. Nhưng khi đưa ra lựa chọn ấy, cần nhìn cả mặt sau của câu 'First come, first served'. Ở phía đối diện với lợi ích dành cho người chiếm trước, có những con người phải gánh lỗi của hệ thống được thử nghiệm trước. Họ thường không được ngồi vào bàn đàm phán. Giống như cha mẹ của Minab.

Ta quay lại điều khoản Martens. Các nguyên tắc nhân đạo và mệnh lệnh của lương tri công chúng. Cụm từ cũ này, trong thời đại vũ khí tự động, rốt cuộc đặt ra một câu hỏi. Chúng ta có thể chịu đựng điều gì, và không thể chịu đựng điều gì. Đó là niềm tin rằng ngay cả trước khi điều khoản được ghi vào bộ luật, trong lương tri của nhân loại đã có một ranh giới không được vượt qua. Khi ICRC gọi vũ khí trực tiếp lấy con người làm mục tiêu là 'ranh giới đạo đức không được vượt qua', khi Guterres nói cùng một điều, họ đang đứng đúng tại vị trí mà Martens đã đặt ra vào năm 1899.

Đây cũng là lý do điều khoản Martens trở thành nền tảng cho một hiệp ước mới trong thời đại vũ khí tự động. Nhìn vào những ví dụ lịch sử mà ICRC nêu ra, nhân loại đã từng cấm trước một số loại vũ khí trước khi chúng được dùng rộng rãi. Tuyên bố Saint Petersburg năm 1868 đã cấm

một số loại đạn nổ nhất định, và năm 1995, vũ khí laser gây mù vĩnh viễn mắt người bị cấm. Cả hai đều là những ranh giới được vạch ra trước khi loại vũ khí ấy trở nên phổ biến trên chiến trường. Theo cách nói của ICRC, đó là những thành công hiếm hoi, không phải 'phản ứng sau khi nỗi đau đã rõ ràng', mà là 'ngăn nỗi đau từ trước'. Trường hợp mìn sát thương thì ngược lại. Chỉ sau khi thiệt hại bừa bãi chồng chất hết lần này đến lần khác, lệnh cấm toàn diện mới ra đời. Khi ICRC kêu gọi cấm ngay vũ khí tự động chống con người, họ muốn đi theo con đường của laser gây mù, không phải con đường của mìn sát thương. Vạch ranh giới trước khi thiệt hại chất đống. Điều khoản Martens nói rằng khoảng trống nơi cần vạch ranh giới ấy, khoảng trống mà điều luật chưa lấp đầy, phải được lấp bằng các nguyên tắc nhân đạo.

Nhưng lương tri không phải là luật. Điều khoản Martens là nguyên tắc bảo vệ con người, nhưng không phải là điều khoản đưa người vi phạm nguyên tắc ấy ra trước tòa. Mệnh lệnh của lương tri công chúng có sức nặng, nhưng nó không nói ai phải chịu trách nhiệm và chịu trách nhiệm bằng cách nào khi mệnh lệnh ấy bị phá vỡ. Ở đây, chúng ta gặp lại khoảng trống trong chương 5. Câu nói rằng nếu không có trách nhiệm thì không có răn đe, cũng không có sự đền trả cho nạn nhân. Điều khoản Martens vạch ranh giới. Nhưng câu hỏi ai chịu trách nhiệm khi ranh giới ấy bị vượt qua vẫn đang chờ câu trả lời.

Vậy luật nhân đạo quốc tế có thể làm máy móc dừng lại không. Đó là câu hỏi mà nhan đề chương này đặt ra. Trả lời thẳng thắn, luật đã có sẵn công cụ để dừng lại trong tay. Nguyên tắc phân biệt, nguyên tắc tương xứng, nguyên tắc phòng ngừa. Ba nguyên tắc này được áp dụng bất kể công nghệ là gì. Như ICRC nói, luật nhân đạo quốc tế có tính 'trung lập về công nghệ', nên dù không nhắc đến vũ khí tự động dù chỉ một chữ, nó đã ràng buộc loại vũ khí ấy một cách mạnh mẽ. Vũ khí không thể dự đoán bị cấm vì có tính bừa bãi, còn vũ khí trực tiếp đánh vào con người xung đột với nguyên tắc phân biệt. Luật đã nói rất nhiều rồi.

Vấn đề không phải là luật nói gì, mà là ai áp dụng luật ấy. Cả ba nguyên tắc của luật đều đặt tiền đề trên phán đoán của con người. Khi phán đoán ấy bị rút xuống còn 20 giây, còn 1 phút, khi 1.000 tọa độ đổ xuống trong 24 giờ, các nguyên tắc của luật vẫn ở nguyên đó, nhưng vị trí của người áp dụng chúng lại dần trống đi. Luật có thể dừng lại. Chỉ có điều, để luật dừng được, trong chuỗi quyết định phải có một con người có thời gian để dừng. Người ấy phải nắm thông tin, thẩm quyền và thời gian. Cụm từ kiểm soát có ý nghĩa, điều khoản Martens cũ kỹ, lập luận cấm của ICRC và Guterres, tất cả đều chỉ về đúng vị trí ấy. Vị trí để phán đoán của con người bước vào.

Cuốn sách này đã lần theo quá trình vị trí ấy dần bị bỏ trống. Trên bàn làm việc ở Gaza, trên cánh đồng Ukraine, phía trên ngôi trường ở Iran, trong đêm mất điện ở Caracas. Và ở cuối chương 5, cuốn sách đã nói rằng việc lấp khoảng trống ấy là phần việc của luật. Đi qua chương này, chúng ta đã thấy luật có thể làm việc đó, nhưng vẫn chưa làm được. Hiệp ước đang được thảo luận với mốc thời gian là năm 2026, nhưng quanh bàn ấy có sự ủng hộ của 156 quốc gia

cùng với lực hãm từ một vài cường quốc. Và trong khi cuộc thảo luận ấy tiếp tục, các hệ thống đã được kiểm chứng ở Gaza trở thành hàng hóa, rồi được bán sang những chiến trường khác.

Loại vũ khí mà luật pháp đang cố đuổi kịp ấy, ngay lúc này, vẫn đang được xuất khẩu đến một nơi nào đó. Ở chương tiếp theo, chúng ta sẽ nhìn vào chiến trường đang được xuất khẩu ấy.

# Lời bạt: lethal beta, chiến trường được xuất khẩu

Nhà báo công nghệ Jacob Ward gọi nó là "lethal beta". Một bản beta gây chết người. Trong ngành phần mềm, beta là phiên bản thử nghiệm chưa hoàn thiện. Vẫn còn lỗi, người dùng dùng rồi báo lỗi, và lỗi đó được sửa ở phiên bản sau. Những ứng dụng chúng ta dùng hằng ngày cũng từng là beta. Tin nhắn đôi khi đến chậm, màn hình hiện lên trễ một nhịp, rồi báo cáo lỗi tích lại ở đâu đó, và bản cập nhật sau âm thầm sửa chúng. Beta vốn là như vậy. Đưa một thứ chưa hoàn chỉnh ra đời trước, gom lấy bất tiện của người dùng, rồi tiến dần đến bản hoàn chỉnh.

Nhưng bản beta mà Ward chỉ ra lại giết người. Nó là bản thử nghiệm, nhưng bên báo lỗi lại chết. Muốn báo một lỗi, trước hết người đó phải trở thành đối tượng của chính lỗi ấy. Người bị đưa nhầm vào danh sách không có đường nào để phản đối rằng mình bị đưa nhầm. Người đó đã biến mất rồi. Chủ thể duy nhất có thể gửi báo cáo lỗi lại biến mất vì trở thành chính lỗi đó. Hai chữ của Ward đâm trúng cấu trúc ấy.

Trong beta còn ẩn một lời hứa khác. Lời hứa rằng sẽ có phiên bản sau. Bây giờ chưa hoàn chỉnh, nhưng rồi sẽ tốt hơn. Vì có niềm tin ấy, chúng ta chịu đựng beta. Bản beta gây chết người cũng đi kèm cùng một lời hứa. Tỷ lệ lỗi 10 phần trăm sẽ thành 5 phần trăm ở phiên bản sau, rồi 5 phần trăm lại thành 2 phần trăm. Dữ liệu tốt hơn, xử lý nhanh hơn, tọa độ chính xác hơn. Các con số có thể thật sự đẹp lên. Nhưng cái bẫy nằm ở đây. Con số càng đẹp, con người càng tin. Càng tin, 20 giây thành 10 giây, 10 giây thành 5 giây. Càng chính xác thì càng nhanh, càng nhanh thì khoảng trống để con người chen vào càng hẹp. Đến ngày beta trở thành bản chính thức, con người vẫn ngồi trước màn hình, nhưng không còn phán đoán nữa. Họ chỉ còn là bàn tay bấm nút.

Nghĩ đến nơi cụm từ ấy ra đời, sống lưng lạnh đi. Ở Gaza, một hệ thống tên là Lavender đưa con người vào danh sách. Sĩ quan tình báo nhận danh sách đó và đóng dấu trong 20 giây. Hệ thống ấy được đưa tin là có tỷ lệ lỗi khoảng 10 phần trăm. Nghĩa là trong mười người, có thể có một người hoàn toàn sai. Với phần mềm thông thường, đó là con số đủ để hoãn phát hành. Nếu một ứng dụng thanh toán cứ mười lần lại gửi tiền nhầm sang một tài khoản khác, ứng dụng ấy sẽ không được đưa ra thị trường. Nhưng trên chiến trường, nó đã xuất hiện. Một hệ thống cứ mười lần lại có thể chỉ nhầm một người đã vận hành ở nơi phân định sống chết. Gaza là nơi cho chạy nguyên mẫu.

Nguyên mẫu không dừng lại ở Gaza. Một hệ thống nhắm mục tiêu cùng dòng đã vận hành trên bầu trời Iran năm 2026. Màn hình phán định sống chết của một người hiện lên chỉ trong 60 giây, và một trường học ở Minab vẫn nằm trong danh sách mục tiêu vì một tọa độ cũ. Nếu Gaza là nguyên mẫu, Iran là sân khấu.

Tôi không dùng chữ sân khấu một cách nhẹ tay. Nguyên mẫu được làm trong phòng kín. Sân khấu thì có người xem. Màn hình của chiến dịch Iran chạy trong phòng tác chiến của Mỹ, và kết quả lan qua vệ tinh, qua đường liên lạc, sang màn hình của các nước đồng minh. Một hệ thống chạy trơn tru trở thành thứ để khoe. Nhanh đến mức này, chính xác đến mức này. Một công cụ đã bước lên sân khấu thì tạo ra khán giả. Khán giả rồi sẽ thành người mua.

Sau sân khấu là xuất khẩu. Những hệ thống được lần theo trong cuốn sách này đã được bán cho quân đội và cơ quan tình báo của nhiều nước. Một công cụ đã được kiểm chứng trên một chiến trường sẽ được chuyển sang chiến trường tiếp theo. Trên cánh đồng Ukraine, drone học cách nhận mục tiêu bằng mắt. Ở Gaza, danh sách thay tốc độ của con người bằng tốc độ của máy. Ở Iran, hai thứ ấy nhập vào cùng một màn hình. Thứ đã chạy ở Gaza chạy ở Iran, thứ đã chạy ở Iran sẽ chạy ở một nơi nào đó tiếp theo. Tiếp theo là đâu? Suốt quá trình điều tra, tôi không gặp được ai có câu trả lời cho câu hỏi ấy. Nói cho đúng, tôi có gặp vài người nói rằng họ có câu trả lời. Nhưng câu trả lời của họ đều dừng lại ở ranh giới công ty của họ, đơn vị của họ, hợp đồng của họ. Khi tôi hỏi công cụ ấy sẽ đi xa đến đâu ngoài ranh giới đó, ai cũng có vẻ mặt gần giống nhau. Vẻ mặt nói rằng đó không phải phần việc của tôi.

Ở đây, tôi phải đưa ra một mâu thuẫn mà cuốn sách này đã không thể buông bỏ đến tận cuối.

Ngay cả những người tạo ra và bán công nghệ này cũng đã dừng lại trước một số ranh giới. Dario Amodei, CEO của Anthropic, công ty làm ra AI Claude, dù làm việc với Bộ Quốc phòng Hoa Kỳ, vẫn vạch ra hai giới hạn. Công ty của ông đã làm việc với quân đội sớm hơn bất kỳ công ty AI nào khác. Ông nói chính họ là bên đầu tiên đưa mô hình lên đám mây mật và cũng là bên đầu tiên tạo ra mô hình riêng cho an ninh quốc gia. Tác chiến mạng, hỗ trợ chiến đấu, phân tích tình báo. Mô hình của công ty ấy đã được đặt trong nhiều nơi của quân đội và các cơ quan tình báo. Ông tin rằng phải bảo vệ đất nước khỏi kẻ thù, phải bảo vệ đất nước khỏi các quốc gia độc đoán như Trung Quốc và Nga.

Con người ấy đã nói với Bộ Quốc phòng rằng 98%, 99% những việc các ông định làm là ổn. Nhưng có hai việc thì không được. Trong cuộc phỏng vấn với CBS, ông nói rõ ràng hai giới hạn đó.

Một là giám sát trong nước trên quy mô lớn. Nếu chính phủ mua dữ liệu do các công ty tư nhân thu thập rồi dùng AI để phân tích cùng lúc, hồ sơ của từng công dân có thể được tạo ra, gắn cả vị trí và khuynh hướng chính trị. Ông nói rằng trước AI, việc này không ai làm vì chẳng có mấy giá trị sử dụng, nhưng bây giờ nó đã trở thành có thể. Công nghệ chạy nhanh hơn luật, rồi lách qua khoảng trống của luật. Ông nói điều đó không thể đi cùng các giá trị dân chủ của Hoa Kỳ.

Điều còn lại là vũ khí tự chủ hoàn toàn. Một thứ vũ khí tự bắn mà không có con người can dự. Ông không nói đến vũ khí tự chủ một phần đang được dùng ở Ukraine, mà chỉ thứ vũ khí trong đó con người bị loại hẳn khỏi cò súng. Ông nói thế này. AI ngày nay chưa đáng tin đến mức có

thể giao việc đó. Trong mô hình có một tính bất định căn bản, và bất cứ ai từng làm việc với AI đều biết điều ấy. Chúng ta chưa giải được sự bất ổn đó bằng kỹ thuật. Rồi ông nói rõ một điểm. Ông không phản đối vũ khí tự chủ hoàn toàn về nguyên tắc. Nếu đối thủ có nó, một ngày nào đó chúng ta cũng có thể cần. Nhưng hiện nay độ tin cậy chưa đạt đến mức ấy, và chúng ta vẫn chưa có cuộc đối thoại về giám sát.

Một đoạn trong lời Amodei đọng lại rất lâu. Ông nêu vấn đề trách nhiệm. Hãy tưởng tượng một người, hoặc một nhóm nhỏ, điều khiển 10 triệu drone. Ai là người giữ nút bấm, ai có thể nói không. Ông nói chúng ta chưa từng thật sự có cuộc đối thoại ấy. Người lính con người có lẽ thường, và chuỗi trách nhiệm được dựng lên trên giả định về lẽ thường đó. AI nhắm nhầm người, bắn vào dân thường, hoặc không thấy được phán đoán mà một người lính con người sẽ thấy. Khi chuyện ấy xảy ra, ai chịu trách nhiệm. Nếu rút con người khỏi chuỗi đó, còn lại điều gì.

Đó là lời của người đã vạch ranh giới. Nhưng thực tế lại trôi đi lệch khỏi lời ấy. Claude của cùng công ty được đưa vào hệ thống của Palantir và xuất hiện trên màn hình tác chiến liên quan đến Iran và Venezuela. Người tạo ra nó vạch ranh giới với vũ khí tự chủ hoàn toàn, nhưng mô hình ông tạo ra lại đang chạy ở một góc màn hình chọn mục tiêu. Ranh giới ông vạch là vũ khí tự chủ 'hoàn toàn'. Chừng nào con người còn ngồi trước màn hình và đóng dấu phê chuẩn, ranh giới ấy chưa bị vượt qua. Dù là 20 giây hay 60 giây, chỉ cần con người bấm là được. Ranh giới đã được giữ. Trên danh nghĩa. Khoảng cách giữa việc vạch ranh giới và những gì diễn ra bên trong ranh giới ấy. Khoảng cách đó là chỗ trống mà cuốn sách này đến cuối vẫn chưa lấp được.

Alex Karp của Palantir nói bằng một ngôn ngữ khác. Trong bài viết trên New York Times năm 2023, ông gọi hiện tại là 'khoảnh khắc Oppenheimer'. Đó là khoảnh khắc nhà vật lý làm ra bom hạt nhân, sau Hiroshima, nói rằng họ đã biết tội lỗi của mình, một hiểu biết không thể đánh mất. Karp mượn phép so sánh ấy, nhưng không nói rằng vì thế hãy dừng lại. Ông nói điều ngược hẳn. Ông cho rằng lời kêu gọi dừng lại là sai.

Lập luận của ông là thế này. Kẻ thù sẽ không dừng lại. Họ không thể nào chờ trong khi chúng ta tổ chức những cuộc tranh luận như sân khấu về rủi ro của công nghệ. Họ cứ tiến lên. Nếu xã hội dân chủ tự do muốn thắng, lời kêu gọi đạo đức là không đủ, cần có sức mạnh cứng. Và sức mạnh cứng của thế kỷ này được dựng trên phần mềm. Không thể vì sợ công cụ sắc bén quay lại phía mình mà tránh làm ra nó. Vũ khí AI phải được theo đuổi, mài giũa và kiểm soát bằng quy định. Theo đuổi, mài giũa, kiểm soát. Ba động từ đứng cạnh nhau trong một câu. Không có lời nào bảo dừng lại.

Trong bài viết của Karp có một câu rất thẳng. Phần mềm của Palantir được các cơ quan quốc phòng và tình báo của Mỹ cùng các nước đồng minh dùng để chọn mục tiêu, lập kế hoạch tác chiến và trinh sát vệ tinh. Năng lực hỗ trợ loại bỏ kẻ địch chính là tiền đề tạo nên giá trị của phần mềm ấy. Ông không nói vòng. Thứ chúng tôi làm ra là công cụ giúp việc giết người, và

chính vì thế nó có giá. Ông cũng quở trách các kỹ sư ở Thung lũng Silicon. Những người có thể viết thuật toán đặt quảng cáo tốt hơn, nhưng lại không muốn làm phần mềm cho quân đội Mỹ. Ông viết rằng họ đã quên sự giàu có và đế chế mà họ đang hưởng có được nhờ ai. Dù không trả món nợ ấy, ít ra cũng phải biết mình đang nợ.

Hãy đặt hai người ấy cạnh nhau. Một người vạch ranh giới, một người nói rằng không thể vì sợ công cụ sắc bén mà tránh làm ra nó. Một người khựng lại vì không thể tin cậy, một người thúc tới vì kẻ địch không dừng nên chúng ta cũng không thể dừng. Có lẽ cả hai đều thật lòng với lời mình nói. Cả hai đều nói về lòng yêu nước, cả hai đều nói về những kẻ thù độc tài, cả hai đều nói về dân chủ tự do. Nhưng sản phẩm của hai công ty lại gặp nhau trên cùng một màn hình tác chiến. Dù lời nói khác nhau thế nào, trong thực tế chúng cùng soi vào một tọa độ. Mô hình của người đã vạch ranh giới, và nền tảng của người nói đừng dừng lại, đều nằm đâu đó trên màn hình nhìn xuống Minab.

Giữa hai người ấy, tôi nhớ đến một cảnh ở Venezuela. Đầu năm 2026, cùng một nền tảng đã chạy trong chiến dịch tại Caracas. Rồi có tin cho biết trong nội bộ công ty đã xuất hiện rạn nứt quanh chiến dịch ấy. Một số kỹ sư chỉ muộn màng biết mã mình viết được dùng vào đâu, và họ dao động. Sự do dự mà Karp từng quở trách đã bùng lên ngay trong chính công ty ấy. Nói rằng đừng dừng lại thì dễ. Biết rằng đoạn mã do tay mình viết đã hiện lên trên màn hình cuối cùng của một con người thì không dễ như vậy. Việc vạch ranh giới và việc do dự không chỉ tồn tại bên ngoài công ty. Nó ở ngay trong cùng một tòa nhà, trước cùng một bàn làm việc.

Tôi cứ nhai đi nhai lại cụm từ khoảnh khắc Oppenheimer. Với Oppenheimer, có một sự biết mà ông không thể đánh mất. Ông nói rằng ông đã biết tội lỗi. Trong một bài giảng tại Viện Công nghệ Massachusetts, ông nói những nhà vật lý đã làm ra bom biết tội lỗi, và đó là một sự biết không thể mất đi. Sự biết ấy có chủ thể. Ông đã làm ra nó, và ông đã biết. Người đứng trong phòng thí nghiệm Los Alamos là ông, người nhìn thấy tia chớp đầu tiên tại cuộc thử nghiệm Trinity cũng là ông, người nghe tin Hiroshima cũng là ông. Ở đó có một đôi vai để gánh tội.

Trong khoảnh khắc hiện nay, chủ thể ấy mờ đi. Khi Lavender đưa một cái tên vào danh sách, ai đã giết người đó. Thuật toán tạo ra danh sách, sĩ quan tình báo đóng dấu trong 20 giây, quân đội đã mua hệ thống ấy, công ty bán mô hình cho quân đội đó, hay những người đầu tư vào công ty ấy. Khi trường học ở Minab vẫn nằm trong danh sách mục tiêu vì tọa độ cũ, trách nhiệm không cập nhật tọa độ thuộc về ai. Người nhập tọa độ, hệ thống không kiểm tra dữ liệu đầu vào, hay quy trình đã dạy người ta tin vào hệ thống ấy. Có nhiều ngón tay, nhưng bàn tay bóp cò không tụ lại thành một.

Trong suốt quá trình tìm hiểu, tôi cứ nghĩ mãi vì sao nó trở nên mờ như vậy. Một câu trả lời nằm trong cụm từ thiên lệch tự động hóa. Con người có xu hướng tin câu trả lời do máy đưa ra hơn phán đoán của chính mình. Khi trên màn hình hiện một dấu đỏ, người ta thường tiếp nhận

dấu ấy thay vì hỏi vì sao nó xuất hiện. Hệ thống càng chính xác, niềm tin ấy càng sâu. Với một người phải đóng dấu trong vòng 20 giây, không còn dư địa để nghi ngờ lại phán định mà màn hình đưa ra. Không nghi ngờ có phải lỗi của người ấy không. Hay đó là lỗi của một hệ thống chỉ cho 20 giây và dạy rằng đừng nghi ngờ. Ở đâu đó giữa hai điều ấy, trách nhiệm lại trượt đi một lần nữa.

Một câu trả lời khác nằm trong cụm từ hộp đen. Ngay cả Karp cũng thừa nhận trong bài viết của mình. Các mô hình mới nhất vận hành như thế nào, vì sao vận hành như vậy, điều đó không rõ ràng ngay cả với các nhà khoa học và lập trình viên đã tạo ra chúng. Không gian do hơn 1 nghìn tỷ biến số tạo ra vượt quá khả năng tính toán của đầu óc con người. Vì vậy, khi mô hình ấy đưa một người vào danh sách, không ai có thể giải thích vì sao người đó bị đưa vào. Người tạo ra nó cũng không biết. Người dùng nó cũng không biết. Người gánh chịu lại càng không biết. Một phán định mà không ai biết lý do trở thành phán định cuối cùng của một con người.

Oppenheimer đã biết tội lỗi. Bởi vì có một chủ thể có thể biết. Trước danh sách do Lavender tạo ra, không có chủ thể biết tội lỗi. Vì nó được thiết kế để ai cũng có thể nói rằng mình chỉ làm một việc nhỏ trong phần của mình. Sĩ quan tình báo nói rằng mình đã tin hệ thống. Người tạo ra hệ thống nói rằng quyết định là do con người đưa ra. Con người đã đưa ra quyết định nói rằng thời gian chỉ có 20 giây. Người đặt ra 20 giây nói rằng tốc độ tác chiến đòi hỏi như vậy. Người định ra tốc độ tác chiến nói rằng kẻ địch nhanh nên không còn cách nào khác. Trách nhiệm không biến mất. Nó bị chia nhỏ, phân tán đến mức không đủ nặng để đè xuống bất cứ nơi nào.

Đó chính là chỗ trống. Không phải trống nên thành chỗ trống, mà là chỗ ở giữa đã bị bỏ trống vì mỗi người đều lùi lại một bước. Khoảng cách mỗi người lùi lại không dài. Chỉ một bước. Nhưng khi những bước lùi ấy cộng lại, ở giữa không còn ai. Không còn đôi vai nào biết tội lỗi. Đôi vai mà Oppenheimer từng mang, giờ không còn ở đâu nữa.

Luật pháp đã biết về chỗ trống này. Luật nhân đạo quốc tế nói rằng ngay cả chiến tranh cũng có quy tắc. Hãy phân biệt binh sĩ và thường dân, đừng gây thiệt hại quá mức so với lợi ích quân sự đạt được. Những quy tắc ấy được viết trên tiền đề rằng con người sẽ phán đoán. Đôi mắt để phân biệt, chiếc cân để so sánh, bàn tay để dừng lại. Nhưng khi danh sách do máy tạo ra và phán đoán bị nén vào 20 giây, đôi mắt, chiếc cân và bàn tay ấy đi đâu. Quy tắc vẫn còn nguyên, nhưng chủ thể giữ quy tắc đã bị phân tán. Điều khoản Martens được lập ra vào thế kỷ 19 nói rằng ngay cả trong trường hợp luật chưa ghi rõ, lương tri của nhân loại vẫn bảo vệ con người. Lương tri của nhân loại. Nơi lương tri ấy cư ngụ cũng từng là đôi vai con người. Khi đôi vai bị phân tán, lương tri cũng phân tán.

Khi khép cuốn sách lại, tôi lại nhớ đến một lớp học.

Minab. Một tòa nhà trường học. Trong màn hình nhìn xuống từ vệ tinh, nơi ấy đã là mục tiêu trong một thời gian. Tọa độ mà hệ thống có đã cũ. Tọa độ không được cập nhật, 60 giây trôi

qua trên màn hình, và ai đó đã ra phán định. Tên và tuổi của những đứa trẻ có mặt trong lớp học hôm ấy, tôi chỉ ghi trong phạm vi đã được xác nhận. Tôi sẽ không mô tả thêm. Vì lời hứa của cuốn sách này là không tiêu thụ chúng như những con số.

Trong thời gian tìm hiểu, tôi đã đặt cùng một câu hỏi cho nhiều người. Ai lẽ ra phải biết rằng tọa độ ấy đã cũ. Câu trả lời lần nào cũng lệch sang bên. Cập nhật tọa độ là việc của bộ phận khác, rà soát mục tiêu có giai đoạn tác chiến riêng, hệ thống không nghi ngờ dữ liệu đầu vào. Không ai nói dối. Tất cả đều nói sự thật. Nhưng dù gom tất cả những sự thật ấy lại, vẫn không xuất hiện người nào có thể bảo vệ lớp học đó. Ô việc của mỗi người đều rõ, còn khoảng giữa các ô thì trống. Trường học ở Minab đã rơi đúng vào khoảng trống ấy.

Tôi chỉ muốn để lại một điều. Trong lớp học ấy có những chiếc ghế. Những chiếc ghế nhỏ. Một chiếc trong số đó bị bỏ trống. Ai lẽ ra phải ngồi trước chiếc ghế trống ấy. Đứa trẻ đó sao. Hay là người nào lẽ ra phải biết tọa độ ấy đã cũ. Bàn tay đã phán định trong 60 giây sao. Công ty đã gọi hệ thống ấy là bản beta rồi tung ra sao. Người nói rằng mình đã vạch ranh giới sao. Người nói đừng dừng lại sao.

Tôi không biết phải đặt ai ngồi vào chiếc ghế ấy. Cuốn sách này không đặt câu trả lời ở đó. Vì giả vờ biết câu trả lời có lẽ chính là cách dễ dàng lấp kín chỗ trống ấy. Khi lấp kín, ta không còn nhìn thấy. Không nhìn thấy thì không còn hỏi. Không hỏi thì tọa độ tiếp theo cũng vẫn cũ. Và hệ thống ấy sẽ được xuất khẩu sang sân khấu kế tiếp.

Gaza là nguyên mẫu. Iran là sân khấu. Tiếp theo sẽ là đâu.

Chiếc ghế ấy vẫn còn trống.

# Support This AI Library Book

If this book was useful, you can support future translations, public editions, and open AI knowledge projects through PayPal.

**PayPal**

**https://paypal.me/kimkjj**

Scan the QR code or open the link above.