

KHOA HỌC THÔNG TIN

Khoa học thông tin đã xuất hiện vào năm 1940 từ những công trình trong 3 lĩnh vực khác nhau: vật lý, cơ học thống kê và viễn thông (telecommunications). Mục tiêu của khoa học thông tin là lượng hóa thông tin và biến thông tin thành một khái niệm khoa học chính xác. Khoa học thông tin được áp dụng trong nhiều lĩnh vực khoa học (nhất là trong sinh học vốn đặt ra nhiều thách thức) và kinh tế học. Khoa học thông tin nhờ tính khái quát lớn của nó là một khoa học đầy hấp dẫn đối với các nhà khoa học và các nhà tư tưởng. Tạp chí Khoa học và Công nghệ Việt Nam xin giới thiệu cùng bạn đọc bài báo *La théorie de l'information existe-t-elle?* của Jérôme Segal (Đại học Paris IV Sorbonne, Đại học Vienne) đăng trên Tạp chí *Pour la Science* số tháng 2.2013.



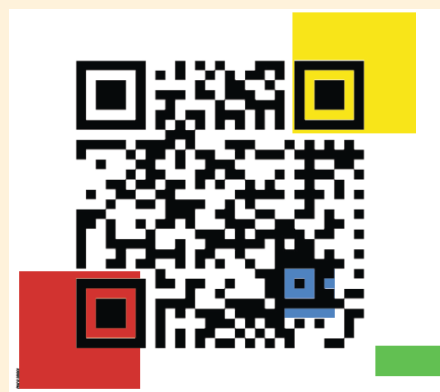
Jérôme Segal

Chúng ta đang sống trong thế giới thông tin (information). Xung quanh chúng ta là một số lượng máy tính khổng lồ ngày đêm thu thập và xử lý thông tin. Quả thật, chúng ta đang chìm ngập trong thông tin. Danh từ thông tin được dùng trong mọi lĩnh vực kinh tế, sinh học, robotics... Nhưng khái niệm thông tin cần được hiểu một cách khoa học, đòi hỏi sự phát sinh của khoa học thông tin. Cần phải lượng hóa (quantifier) khái niệm này. Nhiều nhà khoa học còn cho rằng, *thông tin là một trong những đại lượng vật lý cơ bản nhất, ngang hàng với năng lượng và thời gian*.

Thông tin là một khái niệm với nhiều biến thể không dễ nắm hết được. Sự xuất hiện một lý thuyết khoa học về thông tin chỉ mới được hình thành trong một vài thập kỷ nay và khái niệm thông tin đã được nghiên cứu bởi các nhà khoa học và triết học. Hiện nay, lý thuyết thông tin vẫn còn là một lý thuyết trong quá trình hoàn chỉnh và còn có thể phát triển theo nhiều hướng.

Để hiểu vấn đề này cần phác họa cội nguyên của khoa học thông tin từ năm 1920 và quá trình phát triển của nó. Lý thuyết thông tin có nhiều nguồn gốc và đã thâm nhập vào nhiều lĩnh vực tư duy và hoạt động của con người. Ngược lại, lý thuyết thông tin cũng ảnh hưởng đến sự phát triển của nhiều ngành khoa học (như thông tin génétique, với những gene chứa các dãy bit). Nói chung, một dãy bit có thể có các dạng đồ thị (graphique) khác nhau tùy theo phép mã hóa. Ví dụ, đó là một dãy các vạch thẳng đứng song song trong mã vạch hoặc hình vuông với các pixel đen và trắng như trong mã QR mà nay được sử dụng ngày càng nhiều (hình 1).

Trong quá trình phát triển của khoa học thông tin đã xuất hiện 2 bài báo quan trọng vào năm 1948 là: *Một lý thuyết toán học của truyền thông (communication)* của kỹ sư người Mỹ Claude Shannon và *Điều khiển học (cybernétique)* - điều khiển và truyền thông trong máy móc và sinh vật của nhà toán học người Mỹ Norbert Wiener.

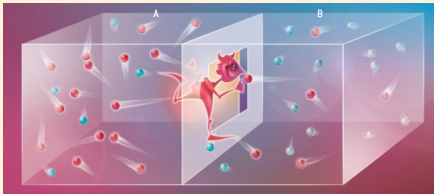


Hình 1: những mã QR (Quick response) sáng chế bởi công ty Nhật Bản Denso-wave là những vạch 2 chiều hiện nay được sử dụng nhiều. Một hình ảnh như thế dễ dàng được thu lại và giải mã nhờ một điện thoại di động có phần mềm thích hợp

Thông tin, một đại lượng khoa học

Vào khoảng giữa 1910 và 1930, trong 3 lĩnh vực rất khác nhau, thông tin được hình thành như một đại lượng khoa học. Trong vật lý, điều này xảy ra gắn liền với thí nghiệm tưởng tượng có tên "*con quỷ của Maxwell - demon de Maxwell*". Con quỷ này là một sinh vật nhỏ tưởng tượng có khả

năng tách những phân tử nhanh ra khỏi những phân tử chậm, như vậy, từ một chất khí đồng nhất nó có thể tạo ra 2 miền với nhiệt độ khác nhau (hình 2).



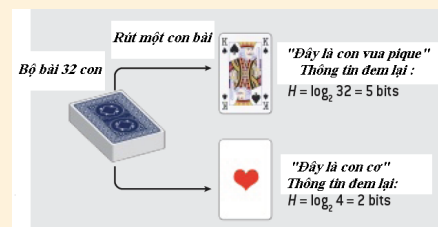
Hình 2: trên hình ta có một chất khí nóng ở A và một chất khí lạnh ở B. Nhờ rằng, nhiệt độ tương ứng với vận tốc trung bình của các phân tử (chính xác hơn nhiệt độ tuyệt đối tỷ lệ trung bình với bình phương của vận tốc các phân tử). Năm 1867, Maxwell đã viết cho một người bạn: "thử tưởng tượng có một sinh vật rất nhỏ có khả năng biết được thông tin về vị trí và vận tốc của tất cả phân tử sinh vật này nhờ đóng và mở một cái cửa để loại các phân tử nhanh từ B sang A và các phân tử chậm từ A sang B. Như vậy, A trở nên nóng hơn và B trở nên lạnh hơn mà không có công nào sản ra ở đây". William Thomson gọi sinh vật này là "con quỷ Maxwell". Thí nghiệm tưởng tượng này đóng vai trò quan trọng trong việc thiết lập mối quan hệ giữa thông tin và entropie

Điều này trái với nguyên lý thứ 2 của nhiệt động học, theo đó, entropie của một hệ cô lập chỉ có thể tăng lên mà thôi (entropie là đại lượng chỉ độ vô trật tự của một hệ. Nói chính xác hơn, *entropie* - *thống kê* đo số lượng các trạng thái vi mô mà hệ có thể tiến đến). Khi bàn đến thí nghiệm tưởng tượng của Maxwell, năm 1912, nhà vật lý người Ba Lan Marjan von Smoluchowski (1872-1917) đã làm rõ vai trò của con quỷ. Con quỷ Maxwell là hình ảnh của một vấn đề lớn: đó là việc thu nhận thông tin liên quan đến sự giảm entropie. Mối liên quan này giữa thông tin và entropie được nghiên cứu tiếp bởi nhà vật lý người Hungary Leo Szilard. Năm 1929, ông đã thử tính xem entropie tăng lên bao nhiêu khi con quỷ Maxwell thu được thông tin về các phân tử

và đã thu được biểu thức $k_B \ln 2$ (k_B là hằng số Boltzmann).

Lượng thông tin, sự bất định và entropie

Muốn xác định lượng thông tin chứa trong một thông điệp có một độ dài nào đó chứa nhiều chữ hay ký tự, Claude Shannon tìm cách lượng hóa tính bất định (tức sự thiếu thông điệp khả dĩ khác có cùng độ dài và cùng sử dụng các chữ hay ký tự như trong bản gốc. Giả sử số thông điệp khả dĩ như vậy bằng N và xác suất chọn thông điệp thứ i là p_i với $p_1 + p_2 + \dots + p_N = 1$, Shannon tìm ra hàm số $H(p_1, \dots, p_N)$ mô tả sự bất định gắn với các khả năng. Shannon chứng minh rằng, hàm H (gọi là entropie thông tin) bằng $-K(p_1 \log p_1 + p_2 \log p_2 + \dots + p_N \log p_N)$, trong đó K là một hằng số. Nếu tất cả $p_i = 0$ ngoại trừ một p nào đó thì sự bất định (incertitude) $= 0$, vì ta chỉ còn một bản gốc. Nếu tất cả p bằng nhau thì $H = K \log(1/p)$ và ta có biểu thức của entropie. Khi thông điệp chỉ chứa 2 số 0 hoặc 1 (bit) thì $p_1 = p_2 = 1/2$ và lúc này $H = K \log 2$. Ví dụ, xét bộ bài gồm 32 con, vậy các p bằng nhau và $= 1/32$, entropie Shannon $= \log_2 32$ vậy có 5 bit vì $32 = 2^5$ (với log cơ số 2) (hình 3).



Hình 3: hàm H của Shannon trong trường hợp bộ bài 32 con

Bằng cách suy rộng khái niệm trên, Shannon còn đề cập đến entropie của một ngôn ngữ. Công thức $-K(p_1 \log p_1 + p_2 \log p_2 + \dots + p_N \log p_N)$ rất giống công thức entropie trong cơ học thống kê của Ludwig Boltzmann và Josiah Gibbs, chỉ cần

lấy $\log$ neperien và cho K là hằng số Boltzmann. Đại lượng p_i chính là xác suất của trạng thái vi mô i của hệ. Sự đồng nhất entropie - thông tin của Shannon với entropie của Boltzmann - Gibbs đã được kiểm nghiệm bởi nhiều nhà khoa học.

Từ thống kê đến thông tin

Một lĩnh vực khác mà ở đây xuất hiện định nghĩa khoa học của thông tin là lĩnh vực thống kê. Từ năm 1922, nhà thống kê người Anh Ronald Fisher đang làm việc tại Trạm nông học Rothamsted đã tìm cách định nghĩa lượng thông tin trên cơ sở những gói dữ liệu để thu được thông tin về những thông số (ví dụ, giá trị trung bình) đặc trưng của định luật phân bố thống kê của dữ liệu. Ông cho rằng, mục tiêu của phương pháp thống kê là sự rút gọn các dữ liệu. Phải thay một đồng dữ liệu bằng một số ít đại lượng đại diện trung thực đồng dữ liệu đó, nói cách khác tìm một số ít đại lượng chứa tối đa thông tin cốt yếu nằm trong số dữ liệu ban đầu. Ví dụ, giả sử độ dài x của các gié lúa về nguyên tắc tuân theo định luật phân bố Gauss (hình cái chuông), độ dài x giảm theo hàm $(x - x_{tb})^2$, trong đó x_{tb} là độ dài trung bình.

Một trong những bài toán trung tâm của lý thuyết thống kê là xác định, xuất phát từ gói dữ liệu mà ta có, những thông số đặc trưng của định luật thống kê tương ứng (như x_{tb} và $(x - x_{tb})^2$ trong ví dụ trên). Mục đích của Fisher là đối với mỗi gói dữ liệu ta phải xác định được "thông tin" để biết được các thông số đó. Năm 1925, Fisher thành công trong việc lượng hóa thông tin nằm trong một gói dữ liệu thống kê. Tiếp sau đó, ông tìm mối liên quan giữa sự tăng entropie với sự mất thông tin và cho rằng "trong những quá trình không thuận nghịch người ta không thể khôi phục được những dữ liệu ban đầu, ở đây có sự mất thông tin".

Một con đường khác dẫn đến

khái niệm khoa học về thông tin là viễn thông (telecommunication). Khi đi tìm cách tăng dung lượng truyền của những đường dây bưu điện, 2 kỹ sư của Phòng thí nghiệm Bell (Mỹ) là Harry Nyquist và Ralph Hartley đã xây dựng những yếu tố đầu tiên của một lý thuyết toán học của truyền thông (communication).

Toán học hóa truyền thông

Năm 1924, Nyquist đã thiết lập rằng, vận tốc truyền thông tin tỷ lệ với với log của số S của các tín hiệu (signaux) hay các ký tự (symboles) cơ bản đã dùng ($S=2$ trong trường hợp mã nhị phân - code binaire). Năm 1928, Hartley định nghĩa lượng thông tin chứa trong một dãy N ký tự (symboles) cơ bản là bằng $N \log S$. Các dòng ý tưởng trên của Nyquist, Hartley và nhiều kỹ sư, nhà khoa học khác cùng với các công trình của Shannon - một nhà toán học - kỹ sư cũng ở Phòng thí nghiệm Bell dẫn đến sự thành hình của “một lý thuyết thông tin”. Năm 1945, Shannon đã viết một tài liệu quan trọng “*Lý thuyết toán học của mật mã*” (tài liệu này được giữ bí mật cho đến năm 1957 mới công bố). Trong công trình này, ông có trực giác so sánh những nhiễu loạn của tín hiệu gây nên bởi mã hóa với “tiếng ồn” vốn đang gây nhiễu các đường dây viễn thông (telecommunications) mà các kỹ sư muốn loại đi.

Lý thuyết thông tin và điều khiển học (cybernetique)

Năm 1975, các nhà vật lý, hóa học và sinh học đã có cuộc hội thảo về vấn đề “Sự thống nhất của khoa học và những giá trị tuyệt đối” để bàn luận xung quanh khái niệm khoa học của thông tin. Nhà sinh học Gunther Stent (Đại học California ở Berkeley - Mỹ) cho rằng, chuyển tải bằng các tín hiệu radio dây các nucleotide của một con mèo đến một nền văn minh ngoài trái đất tương đương với

việc viễn tải chính con mèo. Người ta còn cho rằng, trong tương lai ta có thể chuyển tải nhân cách và cá tính của một người già sang nhân bản trẻ tuổi của chính người già này. Lý thuyết thông tin kết hợp với điều khiển học (cybernetique) có thể dẫn đến nhiều kết quả và nhiều áp dụng quan trọng trong nhiều lĩnh vực khoa học và công nghệ.

Thông tin và những quá trình liên hệ ngược (feedback)

Sự đa ngành gắn liền với chiến tranh đã dẫn đến một lý thuyết tổng quát về điều khiển mà vào năm 1947, Wiener (Đại học Massachusetts, Mỹ) đặt tên là *điều khiển học (cybernetique)*. Lý thuyết này ứng dụng cho mọi hệ thống từ máy móc đến các hệ sinh học, các cơ thể sống. Một yếu tố quan trọng trong điều khiển học là khái niệm liên hệ ngược (feedback - rétroaction).

Các liên hệ ngược rất quan trọng (ví dụ trong cơ học, việc điều khiển áp suất trong máy hơi nước; trong sinh học, việc điều khiển nhiệt độ và huyết áp đều gắn liền với feedback). Từ đó, các nhà điều khiển học xây dựng một lý thuyết trừu tượng của các hệ điều khiển. Ngoài ra, còn phải kể đến các nhà toán học như Neumann, các nhà sinh học như Max Delbruck (là cha đẻ của sinh học phân tử) và xa nữa là các nhà nhân loại học (anthropologues) như Margaret Mead hay Gregory Bateson. Nhiều nhà khoa học muốn phát triển khoa học thông tin thành một khoa học cơ bản liên quan đến mọi lĩnh vực của nhận thức (savoir).

Vật chất, không gian, thời gian và thông tin

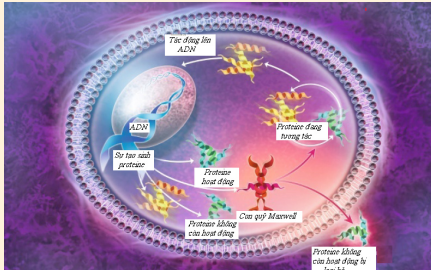
Cách đây hơn 20 năm, Rolf Landauer (1927-1999) một nhà vật lý đã từng làm việc cho Hãng IBM (Mỹ) khẳng định rằng “thông tin là vật lý” và có thể đứng ngang hàng với những yếu tố cơ bản của

tự nhiên như vật chất, năng lượng, không gian và thời gian. Tại Pháp, nhà vật lý năng lượng cao Gilles Cohen - Tannoudji đề nghị đưa vào một hằng số cơ bản thứ 4, hằng số “thông tin” (tổ hợp giữa hằng số Boltzmann và hằng số Planck).

Một quan điểm cũng rất cơ bản là của Vlatko Vedral, người Serbe, Đại học Oxford và Singapore về hướng nghiên cứu *thông tin lượng tử (information quantique)*. Theo nhà vật lý này thì không phải năng lượng, vật chất, không gian hay thời gian mà chính thông tin mới giúp chúng ta hiểu được vũ trụ. Trong quan điểm đó, vũ trụ giống như một máy tính khổng lồ.

Trong những năm 50, lý thuyết thông tin đã cung cấp một ngôn ngữ chung cho các nhà sinh học trong các nghiên cứu nhằm thiết lập mã gen (code genetique) sau phát hiện cấu trúc xoắn đôi của ADN. Nhờ kết hợp những ý tưởng từ điều khiển học (cybernetique) về *tự tổ chức*, các nhà sinh học đã tiến đến việc giải thích sự sống. Hiện nay, cùng với sự đột sinh của *sinh học tổng hợp (biologie de synthèse)*, nhiều nhà sinh học có thể thay đổi sâu sắc các tế bào sống, người ta chứng kiến một luồng ý tưởng mới của lý thuyết thông tin có khả năng dẫn đến nhiều hiện tượng mới. Muốn hiểu một tế bào hoạt động ra sao, các nhà sinh học như Antoine Danchin - một nhà nghiên cứu bộ gen (genomique) vi khuẩn (bacterie) ở Pháp và là người sáng lập Công ty Societe Ambiotics cho rằng, những sinh vật sống là những hệ thu nhận thông tin và xử lý thông tin, ở đây thông tin được định nghĩa một cách rộng rãi như một đại lượng chuyển giao trong quá trình đo đạc, với những mối liên hệ (correlations) được tạo ra giữa hệ bị đo và hệ thực hiện phép đo. Theo A. Danchin, cần cầu cứu đến *con quỷ Maxwell* để hiểu vấn đề: nhờ đâu một tế bào thu nhận thông tin có khả năng phản kháng lại (contrer) những hiệu ứng già hóa và

sau đó tự tái tạo? Vì rằng một tế bào có thể chứa những con quỷ phân tử có khả năng phân biệt các protein bình thường với những protein không còn hoạt động nữa và loại những tế bào hoạt động (hình 4).



Hình 4: sự hoạt động của một sinh thể trông giống như sự hoạt động của con quỷ Maxwell. Một tế bào tạo ra protein nhờ ADN, trong đó một số protein là sai hỏng và cần phải loại bỏ. Một con quỷ Maxwell phân tử vốn cũng được sinh ra bởi tế bào ADN đã thực hiện việc lựa chọn đó. Những thực thể phân tử đóng vai trò quỷ Maxwell cần được phát hiện

Xóa thông tin đòi hỏi có năng lượng

Để hiểu vấn đề này, chúng ta cần đến các công trình của Landauer năm 1960 về entropie + thông tin. Lúc thu nhận thông tin, con quỷ sinh học Maxwell phải thay đổi trạng thái để hoạt động. Quá trình này gắn liền với entropie, do đó với năng lượng. Quả thật, Landauer đã chứng minh rằng, những phép toán logic thuận nghịch như phép toán NON (làm đảo nghịch trị số của một bit thông tin chuyển 0 thành 1 hay 1 thành 0) có thể xảy ra không cần năng lượng nhưng một phép toán không thuận nghịch như phép toán xóa (effacement) đòi hỏi một năng lượng tối thiểu = $k_B T \ln 2$ để làm mất một bit thông tin, trong đó T = nhiệt độ tuyệt đối. Sự tiêu thụ năng lượng này theo Danchin giúp ta nhận dạng được các tế bào quỷ Maxwell. Nếu chúng tồn tại chúng phải được mã hóa bởi genome và đặc biệt hơn bởi "paléome", một tập hợp gen có khả năng mã hóa các chức năng quan

trọng của sự sống.

Như vậy, nếu có một lượng năng lượng bị tiêu hao một cách bất ngờ thì điều này sẽ làm bộc lộ sự có mặt các tế bào quỷ Maxwell. Một tiêu hao năng lượng đáng ngờ đó có thể là nguyên nhân của một cơ chế giúp kiểm tra sự chuyển ARN thành protein, điều này đã được quan sát năm 1970. Ngoài ra, còn những con đường khác để nhận dạng những tế bào quỷ Maxwell. Ví dụ, một họ các protein có tên là *septine* có nhiệm vụ giam giữ các protein già cỗi trong các tế bào men (levure), các protein già này bị giam giữ trong các tế bào mẹ và không được chuyển sang các tế bào con. Các septine này có thể là những con quỷ Maxwell. Nhận dạng các con quỷ Maxwell trong hoạt động của tế bào và tìm xem chúng đã hoạt động như thế nào đối với A. Danchin và các nhà sinh học khác là một công việc quan trọng. Nghiên cứu những vấn đề này còn có ích cho những câu hỏi liệu thông tin có tuân theo nguyên lý bảo toàn như đối với năng lượng hay không?

Xung quanh khái niệm thông tin, người ta còn có thể thiết lập nhiều mối liên hệ quan trọng giữa các lĩnh vực khác nhau. Có thể kể đến lý thuyết cấu trúc (structuralism) phát sinh giữa nhà ngôn ngữ học (linguiste) Nga Roman Jakobson và nhà nhân loại học (anthropologue) Pháp Claude Lévi - Strass. Ngôn ngữ chung của họ dựa trên định nghĩa của khái niệm thông tin.

Một lý thuyết hứa hẹn nhiều sự tốt đẹp trước mắt

Trong giai đoạn được gọi là của *điều khiển học thứ hai* hay *điều khiển học cấp hai*, các hệ tự điều khiển (autoregules) có thể tạo nên trật tự (autopoiesis), lý thuyết thông tin bắt đầu có nhiều ứng dụng quan trọng trong các hệ kinh tế học. Thị trường có thể tự điều chỉnh không cần đến sự can thiệp của chính phủ và các

nhà tư tưởng kinh tế như Friedrich Hayek (Áo) đã đưa ra khái niệm *kinh tế tự do* (conception libérale de l'économie). Sau đó, khái niệm được suy rộng đến những lĩnh vực không hàng hóa và dẫn đến sự xuất hiện của chủ nghĩa neoliberalism (một biến dạng của liberalism trong đó có sự can thiệp có giới hạn của nhà nước theo trường phái Chicago).

Cần phải nói thêm rằng, chính trong các lĩnh vực khoa học và công nghệ chính xác mà khái niệm thông tin phát huy nhiều hiệu quả nhất. Người ta có thể nói đến thiên văn học các lỗ đen, ở đây số phận của thông tin và entropie bị nuốt bởi lỗ đen vẫn còn là những vấn đề lớn tuy đã được khai phá song vẫn còn bỏ ngỏ. Các ứng dụng thiết thân của thông tin là viễn thông (telecommunication) và tin học (informatics). Tại Vienne (Áo), nhóm nghiên cứu dưới sự lãnh đạo của Anton Zeilinger đã thực hiện sự viễn tải (téléportation) những photon trong trạng thái liên đới lượng tử (quantum entanglement) trên một khoảng cách lớn hơn 140 km. Lý thuyết thông tin đã qua một lịch sử dài hơn 65 năm và trong tương lai gần có thể gặt hái nhiều kết quả quan trọng bất ngờ khác ■

CC. biên dịch

Tài liệu tham khảo

- [1] P.M. Binder et A. danchin, *Life's demons: information and order in biology*, EMBO Reports, vol. 12(6), pp. 495-499, 2011.
- [2] V. Vedral, *Decoding reality - The Universe as quantum Information*, Oxford University Press, 2010.
- [3] M. Tricot, *Le moment cybernétique - La constitution de la notion d'information, champ - Vallon*, 2008.
- [4] G. Cohen - Tannoudji, *Les constantes universelles*, Hachette, 1996.
- [5] H. Yockey, *Information Theory and Molecular biology*, Cambridge University Press, 1992.