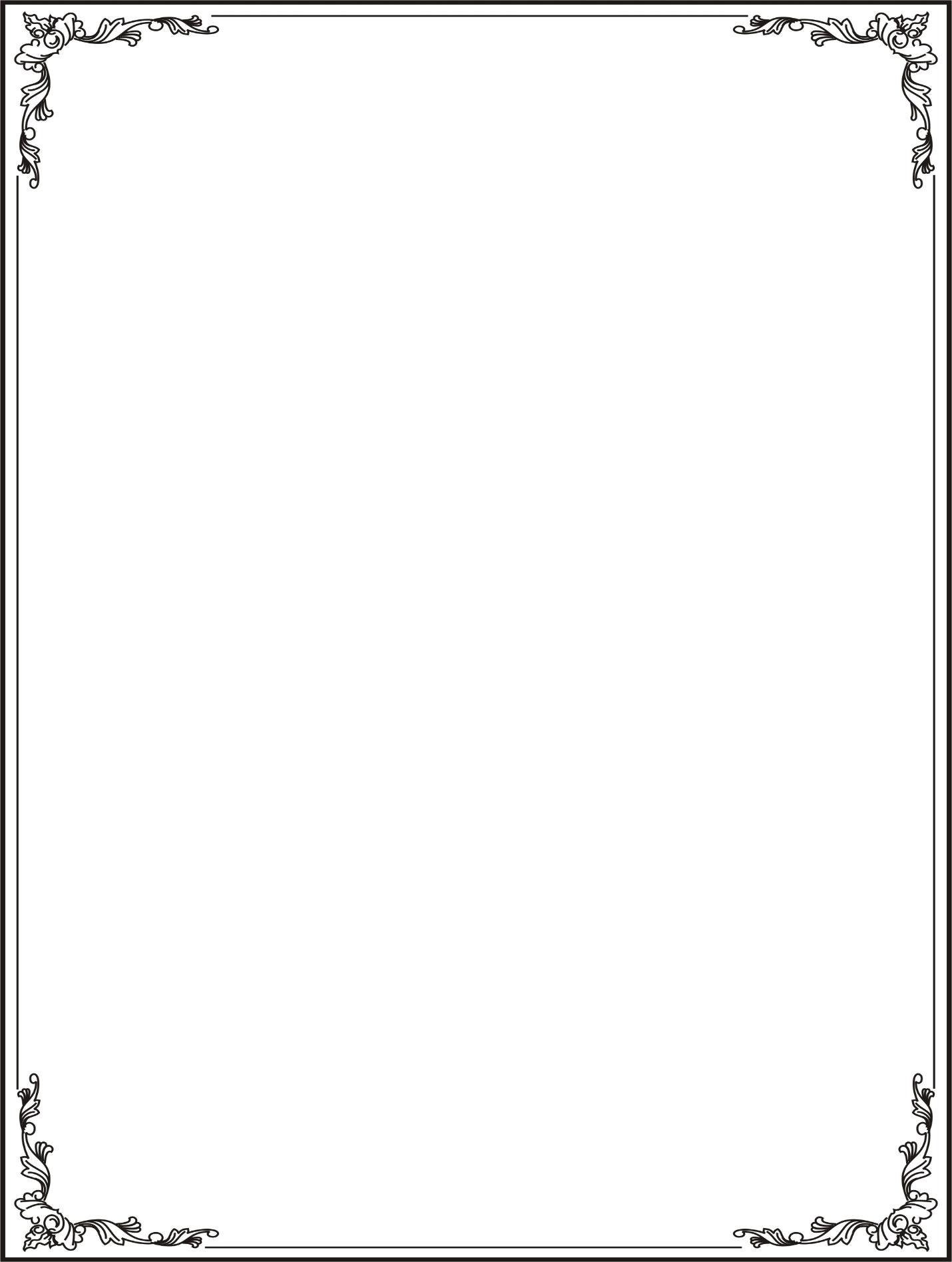
**ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----



**Báo Cáo Bài Tập Lớn**

**Môn Trí Tuệ Nhân Tạo**

***Đề tài* :** Tìm ứng dụng cây quyết định dự báo thu nhập

Giáo viên hướng dẫn: Trần Hùng Cường

Nhóm sinh viên thực hiện: Nhóm 1

Dương Văn Hùng 2019606771

Nguyễn Khắc Huy 2019606088

Nguyễn Văn Hiệp 2019605676

Cao Văn Trung 2019606931

**Hà Nội,Năm 2021**

**Mục Lục**

[**Chương 1.Tổng quan về trí tuệ nhân tạo** 3](#_Toc89377316)

[**1.Khái niệm về trí tuệ nhân tạo** 3](#_Toc89377317)

[Hình 1.1. Phép thử Turing. 4](#_Toc89377318)

[Hình 1.2. Mô hình tác nhân thông minh 5](#_Toc89377319)

[Hình 1.3. Người máy ASIMO đưa đồ uống cho khách theo yêu cầu 5](#_Toc89377320)

[**2.Vai trò của trí tuệ nhân tạo** 5](#_Toc89377321)

[**3. Các kĩ thuật cơ bản trong trí tuệ nhân tạo** 6](#_Toc89377322)

[**4. Lịch sử phát triển của trí tuệ nhân tạo** 7](#_Toc89377323)

[Hình 1.4. Mô hình hệ chuyên gia 9](#_Toc89377324)

[**5. Các thành phần trong hệ thống của trí tuệ nhân tạo** 10](#_Toc89377325)

[**6. Các lĩnh vực nghiên cứu và ứng dụng cơ bản** 10](#_Toc89377326)

[**CHƯƠNG 2: CÂY QUYẾT ĐỊNH** 13](#_Toc89377327)

[**1. CÂY QUYẾT ĐỊNH** 13](#_Toc89377328)

[**2. THUẬT TOÁN TẠO CÂY QUYẾT ĐỊNH** 13](#_Toc89377329)

[**3. VÍ DỤ MINH HOẠ** 15](#_Toc89377330)

# **Chương 1.Tổng quan về trí tuệ nhân tạo**

## **1.Khái niệm về trí tuệ nhân tạo**

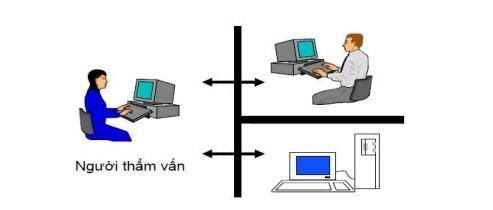
     Trong lĩnh vực Công nghệ thông tin, Trí tuệ nhân tạo (TTNT) cũng có thể hiểu là “thông minh nhân tạo”, tức là sự thông minh của máy móc do con người tạo ra, đặc biệt tạo ra cho máy tính, robot, hay các máy móc có các thành phần tính toán điện tử. TTNT là một ngành mới, nhưng phát triển rất mạnh mẽ và đem lại nhiều kết quả to lớn. Mùa hè 1956, tại hội thảo ở Darmouth John McCarthy đã đưa ra thuật ngữ trí tuệ nhân tạo (Artificial Intelligence - AI). Mốc thời gian này được xem là thời điểm ra đời thực sự của lĩnh vực nghiên cứu TTNT.

     TTNT là một lĩnh vực nghiên cứu của khoa học máy tính và khoa học tính toán nói chung. Có nhiều quan điểm khác nhau về TTNT. Do đó có nhiều định nghĩa khác nhau về lĩnh vực này. Sau đây là một số định nghĩa [3]:

* “Sự nghiên cứu các năng lực trí tuệ thông qua việc sử dụng các mô hình tính toán” (Charniak và McDormott, 1985).
* “Nghệ thuật tạo ra các máy thực hiện các chức năng đòi hỏi sự thông minh khi được thực hiện bởi con người” (Kurweil, 1990).
* “Lĩnh vực nghiên cứu tìm cách giải thích và mô phỏng các hành vi thông minh trong thuật ngữ các quá trình tính toán” (Schalkoff, 1990).
* “Sự nghiên cứu các tính toán để có thể nhận thức, lập luận và hành động” (Winston, 1992).
* “Một nhánh của khoa học máy tính liên quan đến sự tự động hóa các hành vi thông minh” (Luger and Stubblefield, 1993).
* “TTNT là sự nghiên cứu thiết kế các tác nhân thông minh” (Poole, Mackworth and Goebel, 1998).

     Trí tuệ nhân tạo là một nhánh của khoa học và công nghệ liên quan đến việc làm cho máy tính có những năng lực của trí tuệ cong người, tiêu biểu như các khả năng biết suy nghĩ và lập luận để giải quyết vấn đề, biết giao tiếp do hiểu ngôn ngữ và tiếng nói, biết học và tự thích nghi,…[2].

     Mong muốn làm cho máy có những khả năng của trí thông minh con người dã có từ nhiều thế kỷ trước, tuy nhiên TTNT chỉ xuất hiện khi con người sang tạo ra máy tính điện tử. Alan Turing – nhà toán học lỗi lạc người Anh, người được xem là cha đẻ của Tin học do đưa ra cách hình thức hóa các khái niệm thuật toán và tính toán trên máy Turing – một mô hình máy trừu tượng mô tả bản chất việc xử lý các ký hiệu hình thức - có đóng góp quan trọng và thú vị cho TTNT vào năm 1950, gọi là phép thử Turing. Theo Turing: “Trí tuệ là những gì có thể đánh giá được thông qua các trắc nghiệm thông minh”



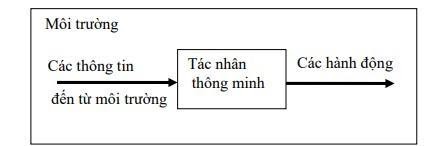
Hình 1.1. Phép thử Turing.

     Phép thử Turing là một cách để trả lời câu hỏi “máy tính có biết nghĩ không?”. Alan Turing đề xuất bộ kiểm thử (Turing test): Trong trắc nghiệm này, một máy tính và một người tham gia trắc nghiệm được đặt vào trong các căn phòng cách biệt với một người thứ hai (người thẩm vấn). Người thẩm vấn không biết được chính xác đối tượng nào là người hay máy tính, và cũng chỉ có thể giao tiếp với hai đối tượng đó thông qua các phương tiện kỹ thuật như một thiết bị soạn thảo văn bản, hay thiết bị đầu cuối. Người thẩm vấn có nhiệm vụ phân biệt người với máy tính bằng cách chỉ dựa trên những câu trả lời của họ đối với những câu hỏi được truyền qua thiết bị liên lạc này. Trong trường hợp nếu người thẩm vấn không thể phân biệt được máy tính với người thì khi đó theo Turing máy tính này có thể được xem là thông minh.

**Khái niệm trí tuệ đưa ra trong từ điển bách khoa toàn thư**:

**Trí tuệ là khả năng**: Phản ứng một cách thích hợp những tình huống mới thông qua hiệu chỉnh hành vi một cách thích đáng. Hiểu rõ những mối liên hệ qua lại của các sự kiện của thế giới bên ngoài nhằm đưa ra những hành động phù hợp đạt tới một mục đích nào đó.

     Hiện nay nhiều nhà nghiên cứu quan niệm rằng, TTNT là lĩnh vực nghiên cứu sự thiết kế các tác nhân thông minh (intellegent agent). Tác nhân thông minh là bất cứ cái gì tồn tại trong môi trường và hành động một cách thông minh.

****

#### Hình 1.2. Mô hình tác nhân thông minh

     Theo M.Minskey: “Trí tuệ nhân tạo mô phỏng bằng máy tính để thí nghiệm một mô hình nào đó “.

     TTNT là một ngành của khoa học máy tính - nghiên cứu xử lý thông tin bằng máy tính, do đó TTNT đặt ra mục tiêu nghiên cứu: làm thế nào thể hiện được các hành vi thông minh bằng thuật toán, rồi nghiên cứu các phương pháp cài đặt các chương trình có thể thực hiện được các hành vi thông minh bằng thuật toán, tiếp theo chúng ta cần chỉ ra tính hiệu quả, tính khả thi của thuật toán thực hiện một nhiệm vụ, và đưa ra các phương pháp cài đặt.

     Mục tiêu của ngành TTNT: Nhằm tạo ra các máy tính có khả năng nhận thức, suy luận và phản ứng. Xây dựng TTNT là tìm cách biểu diễn tri thức và phát hiện tri thức từ các thông tin có sẵn để đưa vào trong máy tính. Để máy tính có các khái niệm nhận thức, suy luận, phản ứng thì ta cần phải cung cấp tri thức cho nó.

****

#### Hình 1.3. Người máy ASIMO đưa đồ uống cho khách theo yêu cầu

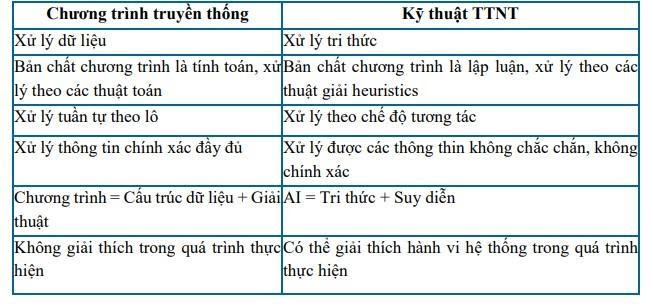
## **2.Vai trò của trí tuệ nhân tạo**

Trí tuệ nhân tạo nghiên cứu kỹ thuật làm cho máy tính có thể “suy nghĩ một cách thông minh” và mô phỏng quá trình suy nghĩ của con người khi đưa ra những quyết định, lời giải. Trên cơ sở đó, ta có thể thiết kế các chương trình cho máy tính để giải quyết bài toán [2].

     Sự ra đời và phát triển của TTNT đã tạo ra một bước nhảy vọt về chất trong kỹ thuật và kỹ nghệ xử lý thông tin. Trí tuệ nhân tạo chính là cơ sở của công nghệ xử lý thông tin mới, độc lập với công nghệ xử lý thông tin truyền thống dựa trên văn bản giấy tờ. Điều này được thể hiện qua các mặt sau:

* Nhờ những công cụ hình thức hoá (các mô hinh logic ngôn ngữ, logic mờ,...), các tri thức thủ tục và tri thức mô tả có thể biểu diễn được trong máy. Do vậy quá trình giải bài toán được thực hiện hiệu quả hơn.
* Mô hình logic ngôn ngữ đã mở rộng khả năng ứng dụng của máy tính trong lĩnh vực đòi hỏi tri thức chuyên gia ở trình độ cao, rất khó như: y học, sinh học, địa lý, tự động hóa.
* Một số phần mềm trí tuệ nhân tạo thể hiện tính thích nghi và tính mềm dẻo đối với các lớp bài toán thuộc nhiều lĩnh vực khác nhau.
* Khi máy tính được trang bị các phần mềm trí tuệ nhân tạo, việc sử dụng mạng sẽ cho phép giải quyết những bài toán cỡ lớn và phân tán.

So sánh kỹ thuật lập trình truyền thống và kỹ thuật xử lý tri thức trong TTNT

****

## **3. Các kĩ thuật cơ bản trong trí tuệ nhân tạo**

Có nhiều kỹ thuật nghiên cứu, phát triển ngành khoa học TTNT. Tuy vậy, các kỹ thuật TTNT thường khá phức tạp khi cài đặt cụ thể, lý do là các kỹ thuật này thiên về xử lý các ký hiệu tượng trưng và đòi hỏi phải sử dụng những tri thức chuyên môn thuộc nhiều lĩnh vực khác nhau. Do vậy, các kỹ thuật TTNT hướng tới khai thác những tri thức về lĩnh vực đang quan tâm được mã hoá trong máy sao cho đạt được mức độ tổng quát, dễ hiểu, dễ diễn đạt thông qua ngôn ngữ chuyên môn gần gũi với ngôn ngữ tự nhiên, dễ khai thác nhằm thu hẹp các khả năng cần xét để đi tới lời giải cuối cùng.

     Các kỹ thuật Trí tuệ nhân tạo cơ bản bao gồm:

- **Lý thuyết giải bài toán và suy diễn thông minh**: Lý thuyết giải bài toán cho phép viết các chương trình giải câu đố, các trò chơi thông qua các suy luận mang tính người.

**- Lý thuyết tìm kiếm may rủi**: Lý thuyết này bao gồm các phương pháp và kỹ thuật tìm kiếm với sự hỗ trợ của thông tin phụ để giải bài toán một cách có hiệu quả.

**- Các ngôn ngữ về TTNT**: Để xử lý các tri thức người ta không chỉ sử dụng các ngôn ngữ lập trình dùng cho các xử lý dữ liệu số, mà cần có ngôn ngữ khác. Các ngôn ngữ chuyên dụng này cho phép lưu trữ và xử lý thông tin ký hiệu. Một số ngôn ngữ được nhiều người biết đến là LISP, PROLOG,...

**- Lý thuyết thể hiện tri thức và hệ chuyên gia**: Trí tuệ nhân tạo là khoa học về thể hiện và sử dụng tri thức. Mạng ngữ nghĩa, logic vị từ, Frame,… là các phương pháp biểu diễn tri thức thông dụng. Việc gắn liền cách thể hiện và sử dụng tri thức là cơ sở hình thành hệ chuyên gia.

**- Lý thuyết nhận dạng và xử lý tiếng nói**: Giai đoạn phát triển đầu của TTNT gắn với lý thuyết nhận dạng. Ứng dụng của phương pháp này trong việc nhận dạng chữ viết, âm thanh,…

- **Người máy**: Cuối những năm 70, người máy trong công nghiệp đã đạt được nhiều tiến bộ. Người máy có bộ phận cảm nhận và các cơ chế hoạt động được nối ghép theo sự điều khiển thông minh. Khoa học về cơ học và TTNT được tích hợp trong khoa học người máy.

- **Tâm lý học xử lý thông tin** : Các kết quả nghiên cứu của tâm lý học giúp Trí tuệ nhân tạo xây dựng các cơ chế trả lời theo hành vi, có ý thức; nó giúp cho việc thực hiện các suy diễn mang tính người.

- Ngoài ra, **xử lý danh sách, kỹ thuật đệ quy, kỹ thuật quay lui và xử lý cú pháp hình thức** là những kỹ thuật cơ bản của tin học truyền thống có liên quan trực tiếp đến TTNT.

## **4. Lịch sử phát triển của trí tuệ nhân tạo**

Lịch sử của TTNT cho thấy ngành khoa học này có nhiều kết quả đáng ghi nhận. Theo các mốc phát triển, người ta thấy TTNT được sinh ra từ những năm 50 với các sự kiện sau:

* Turing được coi là người khai sinh ngành TTNT bởi phát hiện của ông về máy tính có thể lưu trữ chương trình và dữ liệu.
* Tháng 8/1956 J.Mc Carthy, M. Minsky, A. Newell, Shannon. Simon ,… đưa ra khái niêm “trí tuệ nhân tạo”.
* Vào khoảng năm 1960 tại Đại học MIT (Massachussets Institure of Technology) ngôn ngữ LISP ra đời, phù hợp với các nhu cầu xử lý đặc trưng của trí tuệ nhân tạo - đó là ngôn ngữ lập trình đầu tiên dùng cho trí tuệ nhân tạo.
* Thuật ngữ TTNT được dùng đầu tiên vào năm 1961 cũng tại MIT.
* Những năm 60 là giai đoạn lạc quan cao độ về khả năng làm cho máy tính biết suy nghĩ. Trong giai đoạn này người ta đã được chứng kiến máy chơi cờ đầu tiên và các chương trình chứng minh định lý tự động. Cụ thể:

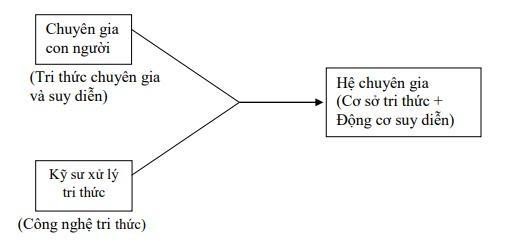
-1961: Chương trình tính tích phân bất định

-1963: Các chương trình heuristics: Chương trình chứng minh các định lý

hình học không gian có tên là “tương tự”, chương trình chơi cờ của Samuel. -1964: Chương trình giải phương trình đại số sơ cấp, chương trình trợ giúp ELIZA (có khả năng làm việc giống như một chuyên gia phân tích tâm lý). --1966: Chương trình phân tích và tổng hợp tiếng nói

-1968: Chương trình điều khiển người máy (Robot) theo đồ án “Mắt – tay”, chương trình học nói.

* Vào những năm 60, do giới hạn khả năng của các thiết bị, bộ nhớ và đặc biệt là yếu tố thời gian thực hiện nên có sự khó khăn trong việc tổng quát hoá các kết quả cụ thể vào trong một chương trình mềm dẻo thông minh.
* Vào những năm 70, máy tính với bộ nhớ lớn và tốc độ tính toán nhanh nhưng các phương pháp tiếp cận TTNT cũ vẫn thất bại do sự bùng nổ tổ hợp trong quá trình tìm kiếm lời giải các bài toán đặt ra.
* Vào cuối những năm 70 một vài kết quả như xử lý ngôn ngữ tự nhiên, biểu diễn tri thức và giải quyết vấn đề. Những kết quả đó đã tạo điều kiện cho sản phẩm thương mại đầu tiên của TTNT ra đời đó là Hệ chuyên gia, được đem áp dụng trong các lĩnh vực khác nhau (Hệ chuyên gia là một phần mềm máy tính chứa các thông tin và tri thức về một lĩnh vực cụ thể nào đó, có khả năng giải quyết những yêu cầu của người sử dụng trong một mức độ nào đó, ở một trình độ như một chuyên gia con người có kinh nghiệm khá lâu năm). Hệ chuyên gia thay thế con người / trợ giúp con người ra quyết định.

****

#### Hình 1.4. Mô hình hệ chuyên gia

* Một sự kiện quan trọng vào những năm 70 là sự ra đời ngôn ngữ Prolog, tương tự LISP nhưng nó có cơ sở dữ liệu đi kèm.
* Vào những năm 80 chứng kiến sự hồi sinh, bùng nổ và thi đua quốc tế trong ngành TTNT. Ý tưởng cơ bản để phát triển TTNT khi này là sự thông minh của máy tính không thể chỉ dựa trên việc suy diễn logic mà phải dựa cả vào (Công nghệ tri thức) (Tri thức chuyên gia và suy diễn) Chuyên gia con người Kỹ sư xử lý tri thức Hệ chuyên gia (Cơ sở tri thức + Động cơ suy diễn) tri thức của con người, và dùng khả năng suy diễn của máy để khai thác tri thức này. Cốt lõi của TTNT có thể diễn giải bởi công thức

                           TTNT = Tri thức + Suy diễn

Thành quả và nỗ lực tiêu biểu trong giai đoạn này là sự phát triển của các hệ chuyên gia. Mỗi hệ chuyên gia về cơ bản gồm hai thành phần: Cơ sở tri thức chứa các tri thức chuyên gia trong một lĩnh vực và một cơ chế suy diễn nhằm vận dụng các hiểu biết này để giải quyết các vấn đề cụ thể với hiệu quả như chính chuyên gia giải quyết. Hai hệ chuyên gia tiêu biểu là DENDRAL và MYCIN. Hệ DENDRAL giúp các nhà nghiên cứu hóa học hữu cơ xác định các phần tử hữu cơ chưa biết dựa trên phân tích phổ của chúng và các tri thức hóa học. MYCIN là hệ chuyên gia y học có cơ sở tri thức khoảng 600 luật.

* Đề án máy tính thế hệ thứ 5 FGCS (Fifth Generation Computer Systems) của Nhật Bản do Bộ Ngoại thương và Công nghiệp phát động. FGCS kéo dài trong 10 năm (1982-1992). Đề án FGCS nhằm làm ra các hệ máy tính có khả năng suy diễn và giao tiếp bằng ngôn ngữ tự nhiên trên nền tính toán song song. Mặc dù cuối cùng được đánh giá là thất bại do không đạt được mục tiêu, nhưng đề án FGCS đã kích thích một cuộc thi đua quốc tế trong giai đoạn hồi sinh của TTNT. Đề án này cũng đặt ra và thách thức nhiều vấn đề cho giới nghiên cứu trên toàn thế giới.
* Những năm 90 cho đến nay, các nghiên cứu nhằm vào cài đặt thành phần thông minh trong các hệ thống thông tin, gọi chung là cài đặt TTNT, làm rõ hơn các ngành của khoa học TTNT và tiến hành các nghiên cứu mới, đặc biệt là nghiên cứu về cơ chế suy lý, về các mô hình tương tác. Các nghiên cứu về AI phân tán, mạng nơron nhân tạo, logic mờ, thuật giải di truyền, khai phá dữ liệu, web ngữ nghĩa, tin sinh học, mạng xã hội,…

## **5. Các thành phần trong hệ thống của trí tuệ nhân tạo**

Hệ thống trí tuệ nhân tạo bao gồm hai thành phần cơ bản đó là biểu diễn tri thức và tìm kiếm tri thức trong miền biểu diễn:

**TTNT = Tri thức + Suy diễn**

     Tri thức của bài toán có thể được phân ra làm ba loại cơ bản đó là tri thức mô tả, tri thức thủ tục và tri thức điều khiển.

Để biểu diễn tri thức người ta sử dụng các phương pháp sau đây:

* Phương pháp biểu diễn nhờ luât
* Phương pháp biểu diễn nhờ mạng ngữ nghĩa
* Phương pháp biểu diễn nhờ bộ ba liên hợp OAV
* Phương pháp biểu diễn nhờ Frame
* Phương pháp biểu diễn nhờ logic vị tư

Sau khi tri thức của bài toán đã được biểu diễn, kỹ thuật trong lĩnh vực trí tuệ nhân tạo là các phương pháp tìm kiếm trong miền đặc trưng tri thức về bài toán đó. Với mỗi cách biểu diễn sẽ có các giải pháp tương ứng. Các vấn đề này sẽ được đề cập trong chương 3.

## **6. Các lĩnh vực nghiên cứu và ứng dụng cơ bản**

Có nhiều nội dung nghiên cứu và phát triển của TTNT, từ cách để máy có thể suy diễn logic và nhận thức, cách ra quyết định và giải quyết vấn đề, cách biểu diễn tri con người trong máy, cách lập kế hoạch hành động, hay biết cách tự học để tạo ra tri thức mới, … đến dịch tự động các ngôn ngữ, tìm kiếm thông tin trên Internet, robot thông minh. Ta nói về một vài lĩnh vực của TTNT trong những năm qua [2].

* Xử lý ngôn ngữ tự nhiên (natural language processing)

*Liệu máy có thể nói được như người?*

     Đây là bài toán tổng hợp tiếng nói, tức việc làm cho máy biết đọc các văn bản thành tiếng người. Có thể hình dung nếu ta đưa cho máy các luật phát âm tiết, bài toán này sẽ là việc áp dụng các luật này vào các âm tiết trong một từ để tạo ra cách đọc từ này. Đã có nhiều hệ thống tạo ra được giọng đọc tự nhiên của con người hoặc đọc giống giọng một người nào đấy, nhất là cho các ngôn ngữ được nghiên cứu nhiều như tiếng Anh.

     Liệu máy có thể nhận biết được tiếng người nói?

     Đây là bài toán nhận dạng tiếng nói, tức việc làm cho máy biết chuyển tiếng nói của người từ microphone thành dãy các từ. Đây là bài toán rất khó, vì âm thanh người nói là liên tục và các âm quyện nối vào nhau, vì mỗi người mỗi giọng,… Với tiếng nói chuẩn, các hệ hiện đại cung mới nhận dạng đúng được khoảng 60%-70%.

     Liệu máy có hiểu được tiếng nói và văn bản của con người?

     Hiểu ngôn ngữ là một đặc trưng tiêu biểu của trí tuệ và việc làm cho máy hiểu được ngôn ngữ là một trong những vấn đề khó nhất của TTNT nói riêng và của CNTT nói chung. Để hiểu nghĩa một câu, máy không chỉ cần biết nghĩa của từng từ, mà trước hết phải biết phân tích được câu này về mặt ngữ pháp. Để làm việc này, máy phải tách câu thành các từ đong lẻ hay cụm từ, nhận viết chúng là các loại từ gì rồi xác định cấu trúc của câu, đoán nghĩa của từng từ và giải nghĩa của câu. Ngôn ngữ thường này trở nên vô cùng khó đối với máy.

     Dịch tự động

     Liên quan đến hiểu ngôn ngữ là dịch tự động từ tiếng này sang tiếng khác. Việc dịch này đòi hỏi máy không chỉ phải hiểu đúng nghĩ a câu tiếng Việt mà còn phải tạo ra được câu tiếng Anh tương ứng.

     Tìm kiếm thông tin trên mạng

     Đây là lĩnh vực có sự chia sẻ nhiều nhất giữa TTNT và Internet, và ngày càng trở nên hết sức quan trọng. Sẽ sớm đến một ngày, mọi sách bào của con người được số hóa và để lên mạng hay các thư viện số cực lớn. Chẳng hạn để tìm các tài liệu có liên quan đến “trí tuệ nhân tạo và ứng dụng trong khoa học”. Với bài toán này có ít nhất hai cách để TTNT đóng góp vào. Một là hệ tìm kiếm các văn bản trong thư viện theo nghĩa này. Hai là hệ tìm kiếm sẽ mô hình các từ “trí tuệ nhân tạo”, “khoa học”, mỗi mô hình là tập hợp các từ khác kèm theo phân bố xác suất của chúng theo quy luật thông kê. Thay vì tìm kiếm trên mạng hay trong thư viện với hai tâp hợp từ khóa, hệ sẽ tìm kiếm với ba tập hợp tư

* Thị giác máy (computer vision): nghiên cứu về việc thu nhận, xử lý, nhận dạng thông tin hình ảnh thành biểu biễn mức cao hơn như các đối tượng xung quanh để máy tính có thể hiểu được.
* Lý thuyết tìm kiếm heuristics: bao gồm các phương pháp và các kỹ thuật tìm kiếm, sử dụng các tri thức đặc biệt nảy sinh từ bản thân lĩnh vực của bài toán cần giải để từ đó nhanh chóng đưa ra kết quả mong muốn. Kỹ thuật cơ bản dựa trên các tri thức Heuristics hay được sử dụng trong thực tiễn là tạo các hàm đánh giá.
* Lý thuyết biểu diễn tri thức và kỹ nghệ xử lý tri thức: Logic mệnh đề, logic vị từ, các hệ sản xuất, biểu diễn bằng Frame, mạng ngữ nghĩa.
* Kỹ thuật suy diễn (inference): Quá trình sinh ra kết luận hoặc sự kiện mới từ những sự kiện và thông tin đã có.
* Học máy (machine learning): Làm tăng hiệu quả giải quyết vấn đề trên dữ liệu và kinh nghiệm đã có.

**Tổng kết**

* Như vậy TTNT là là một lĩnh vực của khoa học và công nghệ nhằm làm cho máy có những khả năng của trí tuệ con người, tiêu biểu như biết suy nghĩ và lập luận để giải quyết vấn đề, biết giao tiếp do hiểu ngôn ngữ tự nhiên và tiếng nói, biết học và tự thích nghi,…
* Sự phát triển của TTNT đã tạo ra một bước nhảy vọt về chất trong kỹ thuật và kỹ nghệ xử lý thông tin. Trí tuệ nhân tạo chính là cơ sở của công nghệ xử lý thông tin mới.
* TTNT có vai trò rất quan trọng trong việc đưa ra lời giải cho các bài toán có không gian tìm kiếm lớn.
* TTNT gồm hai thành phần cơ bản: tri thức và suy diễn

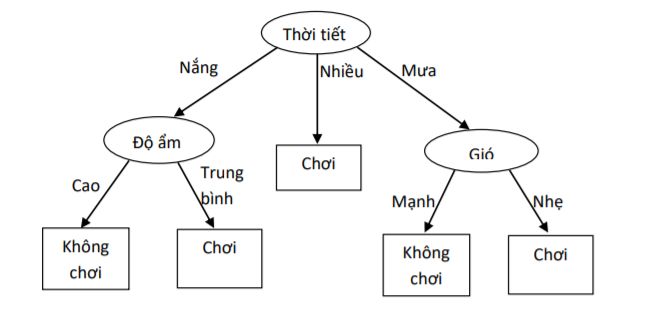
                              AI = Tri thức + Suy diễn

* TTNT được ứng dụng trong hầu hết các lĩnh vực: Kinh tế, địa chất, y học, hóa học,…

# **CHƯƠNG 2: CÂY QUYẾT ĐỊNH**

## **1. CÂY QUYẾT ĐỊNH**

Cây quyết định được dùng để đưa ra tập luật if – then nhằm mục đích dự báo, giúp con người nhận biết về tập dữ liệu. Cây quyết định cho phép phân loại đối tượng tùy thuộc vào các điều kiện tại các nút trong cây, bắt đầu từ gốc cây tới các nút sát lá- Nút xác định phân loại đối tượng. Mỗi nút trong của cây xác định điều kiện đối với thuộc tính mô tả của đối tượng. Mỗi nhánh tương ứng với điều kiện: Nút (thuộc tính) bằng giá trị nào đó. Đối tượng được phân loại nhờ tích hợp các điều kiện bắt đầu từ nút gốc của cây và các thuộc tính mô tả với giá trị của thuộc tính đối tượng.



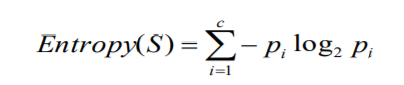
*hình 2.1. Một ví dụ về cây quyết định*

Hình 2.1 là cây quyết định phân loại xem thời tiết như thế nào thì phù hợp với việc chơi tennis

## **2. THUẬT TOÁN TẠO CÂY QUYẾT ĐỊNH**

Xét bảng dữ liệu T = (A, D) trong đó A = {A1, A2,..., An } là tập thuộc tính dẫn xuất, D = {r1, r2, ..., rn} là thuộc tính mục tiêu. Vấn đề đặt ra là trong tập thuộc tính A ta phải chọn thuộc tính nào để phân hoạch? Một trong các phương pháp đó là dựa vào độ lợi thông tin. Hay còn gọi là thuật giải ID3.

Lựa chọn chủ yếu trong giải thuật ID3 là chọn thuộc tính nào để đưa vào mỗi nút trong cây. Ta sẽ chọn thuộc tính phân rã tập mẫu tốt nhất. Thước đo độ tốt của việc chọn lựa thuộc tính là gì? Ta cần xác định một độ đo thống kê, gọi là thông tin thu được, đánh giá từng thuộc tính được chọn tốt như thế nào còn phụ thuộc vào việc phân loại mục tiêu của tập mẫu. ID3 sử dụng thông tin thu được đánh giá để chọn ra thuộc tính cho mỗi bước giữa những thuộc tính ứng viên, trong quá trình phát triển cây.



Để đánh giá chính xác thông tin thu được, dùng Entropy(S): Độ bất định (độ pha trộn/độ hỗn tạp) của S liên quan đến sự phân loại đang xét

 Trong đó pi là xác suất xuất hiện trạng thái i của hệ thống. Theo lý thuyết thông tin: mã có độ dài tối ưu là mã gán –log2 p bits cho thông điệp có xác suất là p. S là một tập huấn luyện.

Nếu gọi p⊕ là xác suất xuất hiện các ví dụ dương trong tập S, pΘ là xác suất xuất hiện các ví dụ âm trong tập S. Entropy đo độ bất định của tập S sẽ là:

Entropy(S) =  -p⊕  log2 p⊕  -  pΘ log2 pΘ

Quy định 0.log 0 = 0

Chẳng hạn với tập S gồm 14 mẫu có chung một vài giá trị logic gồm 9 mẫu dương và 5 mẫu âm. Khi đó đại lượng Entropy của tập S liên quan đến sự phân loại logic này là:

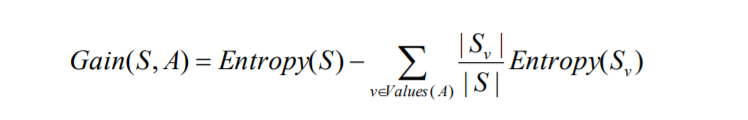
Entropy([9+, 5-]) = - (9/14)log2(9/14) - (5/14)log2(5/14) = 0,940

**Chú ý :**

Đại lượng Entropy = 0 nếu tất cả thành viên của tập S cùng thuộc một lớp (vì nếu tất cả là dương (P+ = 1), do đó P- = 0, Entropy(S) = 1log2 1 0log2 0  0 ).

Đại lượng Entropy(S) = 1 khi tập S chứa tỉ lệ tập mẫu âm và mẫu dương là như nhau. Nếu tập S chứa tập mẫu âm và tập mẫu dương có tỉ lệ P+ khác P- thì Entropy(S) ∈ (0,1).

Dựa trên sự xác định entropy, ta tính Gain(S, A) = Lượng giảm entropy mong đợi qua việc chia các ví dụ theo thuộc tính A



**Ưu/ nhược điểm của thuật toán cây quyết định:**

**Ưu điểm:**

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi với những lợi ích của nó:

* Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
* Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
* Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
* Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
* Có khả năng là việc với dữ liệu lớn

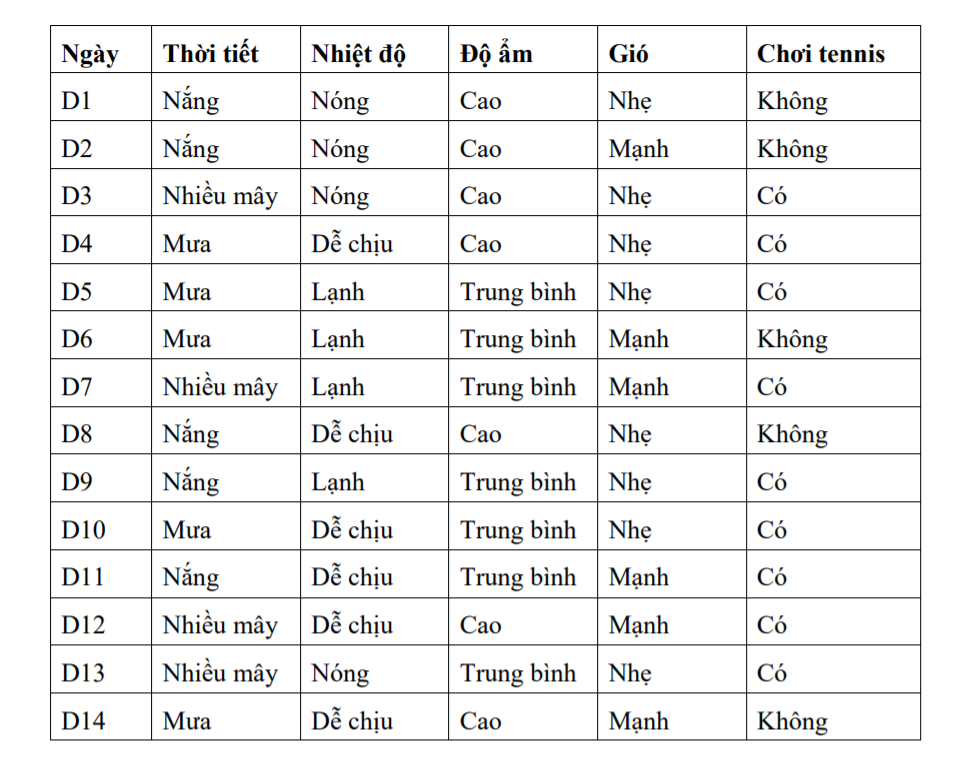
**Nhược điểm:**

Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

* Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
* Cây quyết định hay gặp vấn đề **overfitting** (Overfitting là hiện tượng mô hình ghi nhớ quá tốt dữ liệu huấn luyện và phụ thuộc vào nó, việc này khiến cho mô hình không thể tổng quát hóa các quy luật để hoạt động với dữ liệu chưa từng được chứng kiến).

## **3. VÍ DỤ MINH HOẠ**

Xem xét nhiệm vụ học được đưa ra bởi tập mẫu dưới đây , thuộc tính mục tiêu ở đây là: chơi tennis có giá trị là có hoặc không, giá trị thuộc tính này dự đoán dựa vào các thuộc tính mô tả



Giải quyết bước đầu tiên của giải thuật, tạo nút đỉnh của cây quyết định. Nên đưa thuộc tính nào vào cây đầu tiên? ID3 xác định thông tin thu được cho mỗi thuộc tính ứng cử (thời tiết, nhiệt độ, độ ẩm và gió) sau đó chọn một trong số đó mà có thông tin thu được cao nhất.

Giá trị thông tin thu được cho mỗi thuộc tính là:

Gain(S, thời tiết) = 0,246

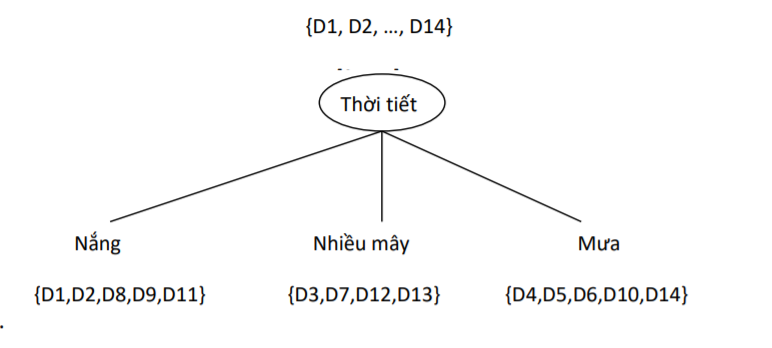
Gain(S, độ ẩm) = 0,151

Gain(S, gió) = 0,048

Gain(S, nhiệt độ) = 0,029

Trong đó tập S là tập mẫu ở bảng trên.

Theo đánh giá thông tin thu được, thuộc tính thời tiết cung cấp dự đoán tốt nhất về thuộc tính mục tiêu “chơi tennis” trên tập mẫu. Do đó, thuộc tính “thời tiết” được chọn là thuộc tính quyết định cho nút gốc, nhánh được tạo ra dưới nút gốc tương ứng với mỗi giá trị của thuộc tính thời tiết (như nắng, mưa, nhiều mây) cùng với tập mẫu sẽ thêm vào mỗi nút con mới.



*Hình 2.2. Cây quyết định sau lần phân hoạch đầu tiên*

Mọi mẫu mà có thời tiết = ‘nhiều mây’ thì là mẫu dương với thuộc tính chơi tennis. Do vậy nút này trở thành nút lá với sự phân loại thuộc tính chơi tennis = ‘Có’. Trái lại với những nút con tương ứng với thời tiết = ‘nắng’ và “thời tiết” = ‘mưa’ có giá trị Entropy ≠ 0 và cây quyết định sẽ phát triển xa hơn dưới những nút này.

Quá trình chọn thuộc tính mới để phân loại tập mẫu lặp lại cho mỗi nút con. Lúc này chỉ sử dụng những mẫu có liên quan tới nút này. Những thuộc tính mô tả có sự kết hợp chặt chẽ hơn trong cây đã được ngăn chặn. Bởi vậy mà bất kì thuộc tính đưa ra nào có thể xuất hiện theo bất kì nhánh nào của cây. Quá trình xử lý còn tiếp cho mỗi nút lá mới cho đến khi hai điều kiện sau thoả mãn: Tập thuộc tính rỗng (mọi thuộc tính đều đã nằm dọc theo những nhánh của cây) hoặc tất cả những mẫu có liên quan với nút lá này có cùng giá trị thuộc tính mục tiêu (giá trị entropy của chúng = 0).