Survey: Artificial Intelligence – Khảo Sát: Trí Tuệ Nhân Tạo

Nguyễn Quản Bá Hồng*

Ngày 17 tháng 5 năm 2025

Tóm tắt nội dung

This text is a part of the series *Some Topics in Advanced STEM & Beyond*:
URL: https://nqbh.github.io/advanced_STEM/.
Latest version:

- *Survey: Artificial Intelligence – Khảo Sát: Trí Tuệ Nhân Tạo.*
  PDF: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/AI/NQBH_AI.pdf.
  TEX: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/AI/NQBH_AI.tex.
- *Lecture Note: Introduction to Artificial Intelligence – Bài Giảng: Nhập Môn Trí Tuệ Nhân Tạo.*
  PDF: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/AI/lecture/NQBH_introduction_AI_lecture.pdf.
  TEX: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/AI/lecture/NQBH_introduction_AI_lecture.tex.
- Codes:
  - C++: https://github.com/NQBH/advanced_STEM_beyond/tree/main/AI/C++.
  - Python: https://github.com/NQBH/advanced_STEM_beyond/tree/main/AI/Python.

Mục lục

1 Basic AI

1.1 [NR21]. Peter Norvig, Stuart Russell. Artificial Intelligence: A Modern Approach

[479 Amazon ratings][4365 Goodreads ratings]

- Amazon review. The long-anticipated revision of *AI: A Modern Approach* explores full breadth & depth of field of AI. 4e brings readers up to date on latest technologies, presents concepts in a more unified manner, & offers new or expanded coverage of ML, DL, transfer learning, multi agent systems, robotics, NLP, causality, probabilistic programming, privacy, fairness, & safe AI.
- Preface. AI is a big field, & this is a big book. Have tried to explore full breadth of field, which encompasses logic, probability, & continuous mathematics; perception, reasoning, learning, & actions; fairness, trust, social good, & safety; & applications that range from microelectronic devices to robotic planetary explorers to online services with billions of users.

  Subtitle of this book is "A Modern Approach". I.e., have chosen to tell story from a current perspective. Synthesize what is now known into a common framework, recasting early work using ideas & terminology that are prevalent today. Apologize to those whose subfields are, as a result, less recognizable.

  - New to 4e. This edition reflects changes of AI since last edition in 2010:
    * Focus more on ML rather than hand-crafted knowledge engineering, due to increased availability of data, computing resources, & new algorithms.
    * DL, probabilistic programming, & multiagent systems receive expanded coverage, each with their own chap.
    * Coverage of natural language understanding, robotics, & computer vision has been revised to reflect impact of DL.

*A scientist- & creative artist wannabe, a mathematics & computer science lecturer of Department of Artificial Intelligence & Data Science (AIDS), School of Technology (SOT), UMT Trường Đại học Quản lý & Công nghệ TP.HCM, Hồ Chí Minh City, Việt Nam.
E-mail: nguyenquanbahong@gmail.com & hong.nguyenquanba@umt.edu.vn. Website: https://nqbh.github.io/. GitHub: https://github.com/NQBH.

* Robotics chap now includes robots that interact with humans & application of reinforcement learning to robotics.
* Previously defined goal of AI as creating systems that try to maximize expected utility, where specific utility information – objective – is supplied by human designers of system. Now we no longer assume: objective is fixed & known by AI system; instead, system may be uncertain about true objectives of humans on whose behalf it operates. It must learn what to maximize & must function appropriately even while uncertain about objective.
* Increase coverage of impact of AI on society, including vital issues of ethics, fairness, trust, & safety.
* Have moved exercises from end of each chap to an online site. This allows us to continuously add to, update, & improve exercises, to meet needs of instructors & to reflect advances in field & in AI-related software tools.
* Overall, about 25% of material in book is brand new. Remaining 75% has been largely rewritten to present a more unified picture of field. 22% of citations of 4e are to works published after 2010.

- Overview of book. Main unifying theme is idea of an *intelligent agent*. Define AI as study of agents that receive percepts from environment & perform actions. Each such agent implements a function that maps percept sequences to actions, & cover different ways to represent these functions, e.g. reactive agents, real-time planners, decision-theoretic systems, & DL systems. Emphasize learning both as a construction method for competent systems & as a way of extending reach of designer into unknown environments. Treat robotics & vision not as independently defined problems, but as occurring in service of achieving goals. Stress importance of task environment in determining appropriate agent design.

  – Chủ đề thống nhất chính là ý tưởng về một *tác nhân thông minh*. Định nghĩa AI là nghiên cứu về các tác nhân tiếp nhận các nhận thức từ môi trường & thực hiện các hành động. Mỗi tác nhân như vậy triển khai một chức năng ánh xạ các chuỗi nhận thức thành các hành động, & bao gồm các cách khác nhau để biểu diễn các chức năng này, ví dụ như các tác nhân phản ứng, các nhà lập kế hoạch thời gian thực, các hệ thống lý thuyết quyết định, & hệ thống DL. Nhấn mạnh việc học vừa là phương pháp xây dựng cho các hệ thống có năng lực & như một cách mở rộng phạm vi của nhà thiết kế vào các môi trường chưa biết. Xử lý robot & tầm nhìn không phải là các vấn đề được xác định độc lập, mà là diễn ra để phục vụ cho việc đạt được các mục tiêu. Nhấn mạnh tầm quan trọng của môi trường nhiệm vụ trong việc xác định thiết kế tác nhân phù hợp.

  Primary aim: convey *ideas* that have emerged over past 70 years of AI research & past 2 millennia of related work. Have tried to avoid excessive formality in presentation of these ideas, while retaining precision. Have included mathematical formulas & pseudocode algorithms to make key ideas concrete; mathematical concepts & notation are described in Appendix A & our pseudocode is described in Appendix B.

  – Mục tiêu chính: truyền đạt *ý tưởng* đã xuất hiện trong hơn 70 năm nghiên cứu AI & 2 thiên niên kỷ công trình liên quan. Đã cố gắng tránh sự trang trọng quá mức trong việc trình bày những ý tưởng này, đồng thời vẫn giữ được độ chính xác. Đã bao gồm các công thức toán học & thuật toán mã giả để làm cho các ý tưởng chính trở nên cụ thể; các khái niệm toán học & ký hiệu được mô tả trong Phụ lục A & mã giả của chúng tôi được mô tả trong Phụ lục B.

  This book is primarily intended for use in an undergraduate course or course sequence. Book has 29 chaps, each requiring about a week's worth of lectures, so working through whole book requires a 2-semester sequence. 1 1-semester course can use selected chaps to suit interests of instructor & students. Book can also be used in a graduate-level course (perhaps with addition of some of primary courses suggested in bibliographical notes), or for self-study or as a reference.

  Only prerequisite is familiarity with basic concepts of CS (algorithms, data structures, complexity) at a sophomore level. Freshman calculus & linear algebra are useful for some of topics.

PART I: AI.

- 1. Introduction. In which we try to explain why we consider AI to be a subject most worthy of study, & in which we try to decide what exactly it is, this being a good thing to decide before embarking.

  Call ourselves *Homo sapiens* – man the wise – because our *intelligence* is so important to us. For thousands of years, have tried to understand *how we think & act* – i.e., how our brain, a mere handful of matter, can perceive, understand, predict, & manipulate a world far larger & more complicated than itself. Field of AI is concerned with not just understanding but also *building* intelligent entities – machines that can compute how to act effectively & safely in a wide variety of novel situations.

  Surveys regularly rank AI as 1 of most interesting & fastest-growing fields, & already generating over a trillion dollars a year in revenue. AI expert Kai-Fu Lee predicts: its impact will be "more than anything in history of mankind". Moreover, intellectual frontiers of AI are wide open. Whereas a student of an older science e.g. physics might feel best ideas have already been discovered by Galileo, Newton, Curie, Einstein, & the rest, AI still has many openings for full-time masterminds.

  AI currently encompasses a huge variety of subfields, ranging from general (learning, reasoning, perception, etc.) to specific, e.g. playing chess, proving mathematical theorems, writing poetry, driving a car, or diagnosing diseases. AI is relevant to any intellectual task; it is truly a universal field.

  * 1.1. What Is AI? Have claimed: AI is interesting, but have not said what it is. Historically, researchers have pursued several different versions of AI. Some have defined intelligence in terms of fidelity to *human* performance, while others prefer an abstract, formal def of intelligence called *rationality* – loosely speaking, doing "right thing". Subject matter itself also varies: some consider intelligence to be a property of internal *thought processes & reasoning*, while others focus on intelligent *behavior*, an external characterization. [In public eye, there is sometimes confusion between terms "AI" & "ML". ML is a subfield of AI that studies ability to improve performance based on experience. Some AI systems use ML methods to achieve competence, but some do not.]

– Đã tuyên bố: AI rất thú vị, nhưng chưa nói rõ nó là gì. Theo truyền thống, các nhà nghiên cứu đã theo đuổi một số phiên bản khác nhau của AI. Một số người đã định nghĩa trí thông minh theo nghĩa là sự trung thành với hiệu suất của *con người*, trong khi những người khác thích một định nghĩa trừu tượng, chính thức về trí thông minh được gọi là *lý trí* – nói một cách rộng rãi, làm "điều đúng đắn". Bản thân chủ đề cũng khác nhau: một số coi trí thông minh là một đặc tính của *quá trình suy nghĩ & lý luận* bên trong, trong khi những người khác tập trung vào *hành vi* thông minh, một đặc điểm bên ngoài. [Trong mắt công chúng, đôi khi có sự nhầm lẫn giữa các thuật ngữ "AI" & "ML". ML là một lĩnh vực phụ của AI nghiên cứu khả năng cải thiện hiệu suất dựa trên kinh nghiệm. Một số hệ thống AI sử dụng các phương pháp ML để đạt được năng lực, nhưng một số thì không.]

From these 2 dimensions – human vs. rational [We are not suggesting that humans are "irrational" in dictionary sense of "deprived of normal mental clarity". We are merely conceding that human decisions are not always mathematically perfect.] & thought vs. behavior – there are 4 possible combinations, & there have been adherents & research programs ∀ 4. Methods used are necessarily different: pursuit of human-like intelligence must be in part an empirical science related to psychology, involving observations & hypotheses about actual human behavior & thought processes; a rationalist approach, on other hand, involves a combination of mathematics & engineering, & connects to statistics, control theory, & economics. Various groups have both disparaged & helped each other. Look at 4 approaches in more detail.

– Từ 2 chiều này – con người so với lý trí [Chúng tôi không ám chỉ rằng con người "phi lý trí" theo nghĩa trong từ điển là "thiếu sự minh mẫn bình thường về mặt tinh thần". Chúng tôi chỉ thừa nhận rằng các quyết định của con người không phải lúc nào cũng hoàn hảo về mặt toán học.] & suy nghĩ so với hành vi – có 4 sự kết hợp có thể, & đã có những người ủng hộ & các chương trình nghiên cứu ∀ 4. Các phương pháp được sử dụng nhất thiết phải khác nhau: việc theo đuổi trí thông minh giống con người phải là một phần khoa học thực nghiệm liên quan đến tâm lý học, bao gồm các quan sát & giả thuyết về hành vi thực tế của con người & các quá trình suy nghĩ; mặt khác, một cách tiếp cận duy lý bao gồm sự kết hợp của toán học & kỹ thuật, & kết nối với thống kê, lý thuyết kiểm soát, & kinh tế học. Nhiều nhóm đã coi thường & giúp đỡ lẫn nhau. Hãy xem xét 4 cách tiếp cận chi tiết hơn.

· 1.1.1. Acting humanly: Turing test approach. *Turing test*, proposed by Alan Turing (1950) was designed as a thought experiment that would sidestep philosophical vagueness of question "Can a machine think?" A computer passes test if a human interrogator, after posing some written questions, cannot tell whether written responses come from a person or from a computer. Chap. 28 discusses details of test & whether a computer would really be intelligent if it passed. For now, note: programming a computer to pass a rigorously applied test provides plenty to work on. Computer would need following capabilities:

1. *natural language processing* to communicate successfully in a human language
2. *knowledge representation* to store what it knows or hears
3. *automated reasoning* to answer questions & to draw new conclusions
4. *ML* to adapt to new circumstances & to detect & extrapolate patterns.

Turing viewed *physical* simulation of a person as unnecessary to demonstrate intelligence. However, other researchers have proposed a *total Turing test*, which requires interaction with objects & people in real world. To pass total Turing test, a robot will need

1. *computer vision* & speech recognition to perceive world
2. *robotics* to manipulate objects & move about.

These 6 disciplines compose most of AI. Yet AI researchers have devoted little effort to passing Turing test, believing: more important to study underlying principles of intelligence. Quest for "artificial flight" succeeded when engineers & investors stopped imitating birds & started using wind tunnels & learning about aerodynamics. Aeronautical engineering texts do not define goal of their fields as making "machines that fly so exactly like pigeons that they can fool even other pigeons".

– 6 chuyên ngành này tạo nên phần lớn AI. Tuy nhiên, các nhà nghiên cứu AI đã dành ít nỗ lực để vượt qua bài kiểm tra Turing, tin rằng: quan trọng hơn là nghiên cứu các nguyên tắc cơ bản của trí thông minh. Nhiệm vụ tìm kiếm "chuyến bay nhân tạo" đã thành công khi các kỹ sư & nhà đầu tư ngừng bắt chước chim & bắt đầu sử dụng đường hầm gió & tìm hiểu về khí động học. Các văn bản về kỹ thuật hàng không không xác định mục tiêu của lĩnh vực này là tạo ra "những cỗ máy bay giống hệt chim bồ câu đến mức chúng có thể đánh lừa cả những con bồ câu khác".

· 1.1.2. Thinking humanly: cognitive modeling approach. To say a program thinks like a human, must know how humans think. Can learn about human though in 3 ways:

1. *introspection* – trying to catch our own thoughts as they go by
2. *psychological experiments* – observing a person in action
3. *brain imaging* – observing brain in action.

Once have a sufficiently precise theory of mind, it becomes possible to express theory as a computer program. If program's input–output behavior matches corresponding human behavior, that is evidence: some of program's mechanisms could also be operating in humans.

– Một khi có một lý thuyết đủ chính xác về tâm trí, có thể diễn đạt lý thuyết như một chương trình máy tính. Nếu hành vi đầu vào-đầu ra của chương trình khớp với hành vi tương ứng của con người, thì đó là bằng chứng: một số cơ chế của chương trình cũng có thể hoạt động ở con người.

E.g., Allen Newell & Herbert Simon, who developed GPS, "General Problem Solver" (Newell & Simon, 1961), were not content merely to have their program solve problems correctly. They were more concerned with comparing sequence & timing of its reasoning steps to those of human subjects solving same problems. Interdisciplinary field of

cognitive science brings together computer models from AI & experimental techniques from psychology to construct precise & testable theories of human mind.
– Ví dụ, Allen Newell & Herbert Simon, người đã phát triển GPS, "General Problem Solver" (Newell & Simon, 1961), không chỉ hài lòng với việc chương trình của họ giải quyết vấn đề một cách chính xác. Họ quan tâm nhiều hơn đến việc so sánh trình tự & thời gian của các bước lý luận của nó với trình tự của con người giải quyết cùng một vấn đề. Lĩnh vực liên ngành của *khoa học nhận thức* tập hợp các mô hình máy tính từ AI & các kỹ thuật thử nghiệm từ tâm lý học để xây dựng các lý thuyết chính xác & có thể kiểm chứng về tâm trí con người.
Cognitive science is a fascinating field in itself, worthy of several textbooks & at least 1 encyclopedia (Wilson & Keil, 1999). Will occasionally comment on similarities or differences between AI techniques & human cognition. Real cognition science, however, is necessary based on experimental investigation of actual humans or animals. Leave that for other books, as assume reader has only a computer for experimentation.
– Khoa học nhận thức là một lĩnh vực hấp dẫn, xứng đáng với một số sách giáo khoa & ít nhất 1 bách khoa toàn thư (Wilson & Keil, 1999). Thỉnh thoảng sẽ bình luận về điểm tương đồng hoặc khác biệt giữa các kỹ thuật AI & nhận thức của con người. Tuy nhiên, khoa học nhận thức thực sự là cần thiết dựa trên nghiên cứu thực nghiệm trên con người hoặc động vật thực tế. Hãy để dành điều đó cho các cuốn sách khác, vì giả sử người đọc chỉ có máy tính để thử nghiệm.
In early days of AI there was often confusion between approaches. An author would argue: an algorithm performs well on a task & therefor a good model of human performance, or vice versa. Modern authors separate 2 kinds of claims; this distinction has allowed both AI & cognitive science to develop more rapidly. 2 fields fertilize each other, most notably in computer vision, which incorporates neurophysiological evidence into computational models. Recently, combination of neuroimaging methods combined with ML techniques for analyzing such data has led to beginnings of a capability to "read minds" – i.e., to ascertain semantic content of a person's inner thoughts. This capability could, in turn, shed further light on how human cognitive works.
– Vào những ngày đầu của AI, thường có sự nhầm lẫn giữa các cách tiếp cận. Một tác giả sẽ lập luận: một thuật toán thực hiện tốt một nhiệm vụ & do đó là một mô hình tốt về hiệu suất của con người, hoặc ngược lại. Các tác giả hiện đại tách biệt 2 loại tuyên bố; sự phân biệt này đã cho phép cả AI & khoa học nhận thức phát triển nhanh hơn. 2 lĩnh vực này hỗ trợ lẫn nhau, đáng chú ý nhất là trong lĩnh vực thị giác máy tính, nơi kết hợp bằng chứng thần kinh sinh lý vào các mô hình tính toán. Gần đây, sự kết hợp của các phương pháp chụp ảnh thần kinh kết hợp với các kỹ thuật ML để phân tích dữ liệu như vậy đã dẫn đến sự khởi đầu của khả năng "đọc suy nghĩ" – tức là xác định nội dung ngữ nghĩa của những suy nghĩ bên trong của một người. Đến lượt mình, khả năng này có thể làm sáng tỏ thêm cách thức hoạt động của nhận thức của con người.

· 1.1.3. Thinking rationally: "law of thought" approach. Greek philosopher Aristotle was 1 of 1st attempt to codify "right thinking" – i.e., irrefutable reasoning processes. His *syllogisms* provided patterns for argument structures that always yielded correct conclusions when given correct premises. Canonical example starts with *Socrates is a man & all men are moral* & concludes *Socrates is mortal*. (This example is probably due to Sextus Empiricus rather than Aristotle). These laws of thought were supposed to govern operation of mind; their study initiated field called *logic*.
Logicians in 19th century developed a precise notation for statements about objects in world & relations among them. (Contrast this with ordinary arithmetic notation, which provides only for statements about *numbers*.) By 1965, programs could, in principle, solve *any* solvable problem described in logical notation. So-called *logicist* tradition within AI hopes to build on such programs to create intelligent systems.
– Các nhà logic học vào thế kỷ 19 đã phát triển một ký hiệu chính xác cho các phát biểu về các đối tượng trong thế giới & các mối quan hệ giữa chúng. (Đối chiếu điều này với ký hiệu số học thông thường, chỉ cung cấp các phát biểu về *số*.) Đến năm 1965, về nguyên tắc, các chương trình có thể giải quyết *bất kỳ* vấn đề có thể giải quyết nào được mô tả bằng ký hiệu logic. Cái gọi là truyền thống logic trong AI hy vọng sẽ xây dựng trên các chương trình như vậy để tạo ra các hệ thống thông minh.
Logic as conventionally understood requires knowledge of world that is *certain* – a condition that, in reality, is seldom achieved. Simply don't know rules of, say, politics or warfare in same way that we know rules of chess or arithmetic. Theory of *probability* fills this gap, allowing rigorous reasoning with uncertain information. In principle, it allows construction of a comprehensive model of rational thought, leading from raw perceptual information to an understanding of how world works to predictions about future. What it does not do, is generate intelligent *behavior*. For that, we need a theory of rational action. Rational thought, by itself, is not enough.
– Logic theo cách hiểu thông thường đòi hỏi kiến thức về thế giới *chắc chắn* – một điều kiện mà trên thực tế hiếm khi đạt được. Đơn giản là không biết các quy tắc của, chẳng hạn, chính trị hay chiến tranh theo cùng cách mà chúng ta biết các quy tắc của cờ vua hay số học. Lý thuyết về *xác suất* lấp đầy khoảng trống này, cho phép lý luận chặt chẽ với thông tin không chắc chắn. Về nguyên tắc, nó cho phép xây dựng một mô hình toàn diện về tư duy hợp lý, dẫn từ thông tin nhận thức thô sơ đến sự hiểu biết về cách thế giới hoạt động để dự đoán về tương lai. Điều mà nó không làm được là tạo ra *hành vi* thông minh. Đối với điều đó, chúng ta cần một lý thuyết về hành động hợp lý. Tư duy hợp lý, tự nó, là không đủ.

· 1.1.4. Acting rationally: rational agent approach. An *agent* is juts sth that acts (*agent* comes from Latin *agere*, to do). Of course, all computer programs do sth, but computer agents are expected to do more: operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, & create & pursue goals. A *rational agent* is one that acts so as to achieve best outcome or, when there is uncertainty, best expected outcome.
– *Hành động hợp lý: phương pháp tiếp cận tác nhân hợp lý.* Một tác nhân chỉ là cái gì đó hành động (tác nhân bắt nguồn từ tiếng Latin agere, nghĩa là làm). Tất nhiên, tất cả các chương trình máy tính đều làm cái gì đó, nhưng các tác

nhân máy tính được kỳ vọng sẽ làm nhiều hơn thế: hoạt động tự chủ, nhận thức môi trường của chúng, tồn tại trong một khoảng thời gian dài, thích nghi với sự thay đổi, & tạo ra & theo đuổi mục tiêu. Một tác nhân hợp lý là tác nhân hành động để đạt được kết quả tốt nhất hoặc, khi có sự không chắc chắn, kết quả mong đợi tốt nhất.
In "laws of thought" approach to AI, emphasis was on correct inferences. Making correct inferences is sometimes *part* of being a rational agent, because 1 way to act rationally: deduce that a given action is best & then to act on that conclusion. On other hand, there are ways of acting rationally that cannot be said to involve inference. E.g., recoiling from a hot stove is a reflex action that is usually more successful than a slower action taken after careful deliberation.
– Trong cách tiếp cận "luật tư duy" đối với AI, trọng tâm là suy luận đúng. Việc đưa ra suy luận đúng đôi khi là *một phần* của việc trở thành một tác nhân lý trí, bởi vì 1 cách để hành động hợp lý: suy ra rằng một hành động nhất định là tốt nhất & sau đó hành động theo kết luận đó. Mặt khác, có những cách hành động hợp lý không thể nói là liên quan đến suy luận. Ví dụ, lùi lại khỏi bếp nóng là một hành động phản xạ thường thành công hơn so với hành động chậm hơn được thực hiện sau khi cân nhắc kỹ lưỡng.
All skills needed for Turing test also allow an agent to act rationally. Knowledge representation & reasoning enable agents to reach good decisions. Need to be able to generate comprehensible sentences in natural language to get by in a complex society. Need learning not only for erudition, but also because it improves our ability to generate effective behavior, especially in circumstances that are new.
– Tất cả các kỹ năng cần thiết cho bài kiểm tra Turing cũng cho phép một tác nhân hành động hợp lý. Biểu diễn kiến thức & lý luận cho phép các tác nhân đưa ra quyết định đúng đắn. Cần có khả năng tạo ra các câu dễ hiểu bằng ngôn ngữ tự nhiên để tồn tại trong một xã hội phức tạp. Cần học không chỉ để có kiến thức uyên bác mà còn vì nó cải thiện khả năng tạo ra hành vi hiệu quả của chúng ta, đặc biệt là trong những hoàn cảnh mới.
Rational-agent approach to AI has 2 advantages over other approaches. 1st, it is more general than "law of thought" approach because correct inference is just 1 of several possible mechanism for achieving rationality. 2nd, more amenable to scientific development. Standard of rationality is mathematically well defined & completely general. Can often work back from this specification to derive agent designs that provably achieve it – sth that is largely impossible if goal: imitate human behavior or thought processes.
– Phương pháp tiếp cận tác nhân hợp lý đối với AI có 2 ưu điểm so với các phương pháp tiếp cận khác. Thứ nhất, nó tổng quát hơn phương pháp tiếp cận "luật tư duy" vì suy luận đúng chỉ là 1 trong số nhiều cơ chế có thể đạt được tính hợp lý. Thứ hai, dễ tiếp thu hơn đối với sự phát triển khoa học. Tiêu chuẩn của tính hợp lý được định nghĩa rõ ràng về mặt toán học & hoàn toàn tổng quát. Thường có thể làm việc ngược lại từ thông số kỹ thuật này để đưa ra các thiết kế tác nhân có thể chứng minh được là đạt được nó – điều mà phần lớn là không thể nếu mục tiêu: bắt chước hành vi hoặc quá trình suy nghĩ của con người.
For these reasons, rational-agent approach to AI has prevailed throughout most of field's history. In early decades, rational agents were built on logical foundations & formed definite plans to achieve specific goals. Later, methods based on probability theory & ML allowed creation of agents that could make decisions under uncertainty to attain best expected outcome. In a nutshell, *AI has focused on study & construction of agents that do right thing.* What counts as right thing is defined by objective that we provide to agent. This general paradigm is so pervasive that we might call it *standard model.* It prevails not only in AI, but also in control theory, where a controller minimizes a cost function; in operations research, where a policy maximizes a sum of rewards; in statistics, where a decision rule minimizes a loss function; & in economics, where a decision maker maximizes utility or some measure of social welfare.
– Vì những lý do này, cách tiếp cận tác nhân hợp lý đối với AI đã chiếm ưu thế trong hầu hết lịch sử của lĩnh vực này. Trong những thập kỷ đầu, các tác nhân hợp lý được xây dựng trên nền tảng logic & hình thành các kế hoạch chắc chắn để đạt được các mục tiêu cụ thể. Sau đó, các phương pháp dựa trên lý thuyết xác suất & ML cho phép tạo ra các tác nhân có thể đưa ra quyết định trong điều kiện không chắc chắn để đạt được kết quả mong đợi tốt nhất. Tóm lại, *AI đã tập trung vào nghiên cứu & xây dựng các tác nhân làm điều đúng đắn.* Những gì được coi là điều đúng đắn được xác định bởi mục tiêu mà chúng ta cung cấp cho tác nhân. Mô hình chung này rất phổ biến đến mức chúng ta có thể gọi nó là *mô hình chuẩn.* Nó không chỉ chiếm ưu thế trong AI mà còn trong lý thuyết điều khiển, trong đó bộ điều khiển giảm thiểu hàm chi phí; trong nghiên cứu hoạt động, trong đó chính sách tối đa hóa tổng phần thưởng; trong thống kê, trong đó quy tắc quyết định giảm thiểu hàm mất mát; & trong kinh tế, trong đó người ra quyết định tối đa hóa tiện ích hoặc một số biện pháp phúc lợi xã hội.
Need to make 1 important refinement to standard model to account for fact that perfect rationality – always taking exactly optimal action – is not feasible in complex environments. Computational demands are just too high. Chaps. 6 & 16 deals with issue of *limited rationality* – acting appropriately when there is not enough time to do all computations one might like. However, perfect rationality often remains a good starting point for theoretical analysis.
– Cần thực hiện 1 cải tiến quan trọng đối với mô hình chuẩn để tính đến thực tế là tính hợp lý hoàn hảo – luôn thực hiện hành động tối ưu chính xác – là không khả thi trong các môi trường phức tạp. Yêu cầu tính toán quá cao. Chương 6 & 16 giải quyết vấn đề về *tính hợp lý hạn chế* – hành động phù hợp khi không có đủ thời gian để thực hiện tất cả các phép tính mà người ta có thể thích. Tuy nhiên, tính hợp lý hoàn hảo thường vẫn là điểm khởi đầu tốt cho phân tích lý thuyết.

· 1.1.5. Beneficial machines. Standard model has been a useful guide for AI research since its inception, but it is probably not right model in long run. Reason: standard model assumes: will supply a fully specified objective to machine.
– Mô hình chuẩn đã là một hướng dẫn hữu ích cho nghiên cứu AI kể từ khi ra đời, nhưng có lẽ về lâu dài, nó không phải là mô hình phù hợp. Lý do: mô hình chuẩn giả định: sẽ cung cấp một mục tiêu được chỉ định đầy đủ cho máy.
For an artificially defined task e.g. chess or shortest-path computation, task comes with an objective built in – so

standard model is applicable. As move into real world, however, it becomes more & more difficult to specify objective completely & correctly. E.g., in designing a self-driving car, one might think: objective is to reach destination safely. But driving along any road incurs a risk of injury due to other errant drivers, equipment failure, etc.; thus, a strict goal of safety requires staying in garage. There is a tradeoff between making progress towards destination & incurring a risk of injury. How should this tradeoff be made? Furthermore, to what extent can we allow car to take actions that would annoy other drivers? How much should car moderate its acceleration, steering, & braking to avoid shaking up passenger? These kinds of questions are difficult to answer a priori. They are particularly problematic in general area of human–robot interaction, of which self-driving car is 1 example.

– Đối với một nhiệm vụ được xác định nhân tạo, ví dụ như cờ vua hoặc tính toán đường đi ngắn nhất, nhiệm vụ đi kèm với một mục tiêu được tích hợp sẵn – do đó, mô hình chuẩn có thể áp dụng được. Tuy nhiên, khi chuyển sang thế giới thực, việc xác định mục tiêu một cách hoàn chỉnh & chính xác trở nên & khó khăn hơn. Ví dụ, khi thiết kế một chiếc xe tự lái, người ta có thể nghĩ: mục tiêu là đến đích an toàn. Nhưng việc lái xe trên bất kỳ con đường nào cũng có nguy cơ bị thương do những người lái xe khác đi sai đường, hỏng thiết bị, v.v.; do đó, mục tiêu an toàn nghiêm ngặt đòi hỏi phải ở trong gara. Có một sự đánh đổi giữa việc tiến về đích & với nguy cơ bị thương. Sự đánh đổi này nên được thực hiện như thế nào? Hơn nữa, chúng ta có thể cho phép xe thực hiện những hành động có thể làm phiền những người lái xe khác ở mức độ nào? Xe nên điều chỉnh gia tốc, đánh lái, & phanh ở mức nào để tránh làm rung chuyển hành khách? Những loại câu hỏi này rất khó trả lời trước. Chúng đặc biệt có vấn đề trong lĩnh vực tương tác giữa con người và rô-bốt nói chung, trong đó xe tự lái là một ví dụ.

Problem of achieving agreement between our true preferences & objective we put into machine is called *value alignment problem*: values or objectives put into machine must be aligned with those of human. In we are developing an AI system in lab or in a simulator – as has been case for most of field's history – there is an easy fix for an incorrectly specified objective: reset system, fix objective, & try again. As field progresses towards increasingly capable intelligent systems that are deployed in rewal world, this approach is no longer viable. A system deployed with an incorrect objective will have negative consequences. Moreover, more intelligent system, more negative consequences.

– Vấn đề đạt được sự đồng thuận giữa sở thích thực sự của chúng ta & mục tiêu mà chúng ta đưa vào máy được gọi là vấn đề căn chỉnh giá trị: các giá trị hoặc mục tiêu đưa vào máy phải phù hợp với mục tiêu hoặc mục tiêu của con người. Khi chúng ta đang phát triển một hệ thống AI trong phòng thí nghiệm hoặc trong trình mô phỏng – như đã từng xảy ra trong hầu hết lịch sử của lĩnh vực này – có một cách khắc phục dễ dàng cho một mục tiêu được chỉ định không chính xác: đặt lại hệ thống, sửa mục tiêu, & thử lại. Khi lĩnh vực này tiến triển theo hướng các hệ thống thông minh ngày càng có khả năng được triển khai trong thế giới mới, cách tiếp cận này không còn khả thi nữa. Một hệ thống được triển khai với mục tiêu không chính xác sẽ có hậu quả tiêu cực. Hơn nữa, hệ thống càng thông minh thì hậu quả tiêu cực càng nhiều.

Returning to apparently unproblematic example of chess, consider what happens if machine is intelligent enough to reason & act beyond confines of chessboard. In that case, it might attempt to increase its chances of winning by such ruses as hypnotizing or blackmailing its opponent or bribing audience to make rustling noises during its opponent's thinking time. [In 1 of 1st books on chess, Ruy Lopez (1561) wrote, "Always place board so sun is in your opponent's eyes."] It might also attempt to hijack additional computing power for itself. *These behaviors are not "unintelligent" or "insane"; they are a logical consequence of defining winning as the sole objective for machine.*

– Quay trở lại ví dụ cờ vua có vẻ không có vấn đề gì, hãy xem xét điều gì xảy ra nếu máy đủ thông minh để lý luận & hành động vượt ra ngoài giới hạn của bàn cờ. Trong trường hợp đó, nó có thể cố gắng tăng cơ hội chiến thắng bằng những mánh khóe như thôi miên hoặc tống tiền đối thủ hoặc hối lộ khán giả tạo ra tiếng sột soạt trong thời gian suy nghĩ của đối thủ. [Trong 1 trong những cuốn sách đầu tiên về cờ vua, Ruy Lopez (1561) đã viết, "Luôn đặt bàn cờ sao cho mặt trời chiếu vào mắt đối thủ."] Nó cũng có thể cố gắng chiếm đoạt thêm sức mạnh tính toán cho chính nó. *Những hành vi này không phải là "không thông minh" hay "điên rồ"; chúng là hệ quả hợp lý của việc xác định chiến thắng là mục tiêu duy nhất của máy.*

Impossible to anticipate all ways in which a machine pursuing a fixed objective might misbehave. There is good reason, then, to think: standard model is inadequate. We don't want machines that are intelligent in sense of pursuing *their* objectives; want them to pursue *our* objectives. If cannot transfer those objectives perfectly to machine, then need a new formulation – one in which machine is pursuing our objectives, but is necessarily *uncertain* as to what they are. When a machine knows that it doesn't know complete objective, it has an incentive to act cautiously, to ask permission, to learn more about our preferences through observation, & to defer to human control. Ultimately, want agents that are *provably beneficial* to humans.

– Không thể lường trước được mọi cách mà một cỗ máy theo đuổi một mục tiêu cố định có thể hoạt động không đúng. Do đó, có lý do chính đáng để nghĩ rằng: mô hình chuẩn là không đủ. Chúng ta không muốn những cỗ máy thông minh theo nghĩa theo đuổi *là mục tiêu của chúng*; muốn chúng theo đuổi *là mục tiêu của chúng ta*. Nếu không thể chuyển những mục tiêu đó một cách hoàn hảo cho cỗ máy, thì cần một công thức mới – công thức mà trong đó cỗ máy đang theo đuổi mục tiêu của chúng ta, nhưng nhất thiết *không chắc chắn* về mục tiêu đó là gì. Khi một cỗ máy biết rằng nó không biết mục tiêu hoàn chỉnh, nó có động cơ để hành động thận trọng, để xin phép, để tìm hiểu thêm về sở thích của chúng ta thông qua quan sát, & để tuân theo sự kiểm soát của con người. Cuối cùng, muốn các tác nhân *có thể chứng minh được là có lợi* cho con người.

∗ 1.2. Foundations of AI. In this sect, provide a brief history of disciplines that contributed ideas, viewpoints, & techniques to AI. Like any history, this one concentrates on a small number of people, events, & ideas & ignores others that also were important. Organize history around a series of questions. Certainly would not wish to give impression that these

questions are only ones the disciplines address or disciplines have all been working toward AI as their ultimate fruition.
– *Nền tảng của AI.* Trong phần này, hãy cung cấp một lịch sử tóm tắt về các ngành đã đóng góp ý tưởng, quan điểm, & kỹ thuật cho AI. Giống như bất kỳ lịch sử nào, phần này tập trung vào một số ít người, sự kiện, & ý tưởng & bỏ qua những người khác cũng quan trọng. Sắp xếp lịch sử xung quanh một loạt các câu hỏi. Chắc chắn không muốn tạo ấn tượng rằng đây chỉ là những câu hỏi mà các ngành giải quyết hoặc tất cả các ngành đều hướng tới AI như là thành quả cuối cùng của họ.

· 1.2.1. Philosophy.

1. Can formal rules be used to draw valid conclusions?
2. How does mind arise from a physical brain?
3. Where does knowledge come from?
4. How does knowledge lead to action?

p. 24+++

◦ 2. Intelligent Agents. In which we discuss nature of agents, perfect or otherwise, diversity of environments, & resulting menagerie of agent types.

– *Các tác nhân thông minh.* Trong đó chúng ta thảo luận về bản chất của các tác nhân, hoàn hảo hay không, sự đa dạng của môi trường, & sự kết hợp của các loại tác nhân.

Chap. 1 identified concept of *rational agents* as central as our approach to AI. In this chap, make this notion more concrete. See: concept of rationality can be applied to a wide variety of agents operating in any imaginable environment. Our plan in this book: use this concept to develop a small set of design principles for building successful agents – systems that can reasonably be called *intelligent*.

– Chương 1 xác định khái niệm về các tác nhân hợp lý là trung tâm trong cách tiếp cận của chúng ta đối với AI. Trong chương này, hãy cụ thể hóa khái niệm này hơn. Xem: khái niệm về tính hợp lý có thể được áp dụng cho nhiều loại tác nhân hoạt động trong bất kỳ môi trường nào có thể tưởng tượng được. Kế hoạch của chúng tôi trong cuốn sách này: sử dụng khái niệm này để phát triển một tập hợp nhỏ các nguyên tắc thiết kế nhằm xây dựng các tác nhân thành công – các hệ thống có thể được gọi một cách hợp lý là thông minh.

Begin by examining agents, environments, & coupling between them. Observation that some agents behave better than others leads naturally to idea of a rational agent – one that behaves as well as possible. How well an agent can behave depends on nature of environment; some environments are more difficult than others. Give a crude categorization of environments & show how properties of an environment influence design of suitable agents for that environment. Describe a number of basic "skeleton" agent designs, which flesh out in rest of book.

– Bắt đầu bằng cách xem xét các tác nhân, môi trường, & sự kết hợp giữa chúng. Quan sát thấy một số tác nhân hành xử tốt hơn những tác nhân khác dẫn đến ý tưởng về một tác nhân hợp lý – một tác nhân hành xử tốt nhất có thể. Mức độ một tác nhân có thể hành xử tốt như thế nào phụ thuộc vào bản chất của môi trường; một số môi trường khó hơn những môi trường khác. Đưa ra một phân loại thô sơ về các môi trường & cho thấy các đặc tính của một môi trường ảnh hưởng đến thiết kế các tác nhân phù hợp cho môi trường đó như thế nào. Mô tả một số thiết kế tác nhân "bộ xương" cơ bản, được trình bày chi tiết trong phần còn lại của cuốn sách.

∗ 2.1. Agents & Environments. An *agent* is anything that can be viewed as perceiving its *environment* through *sensors* & acting upon that environment through *actuators*. This simple idea is illustrated in Fig. 2.1: Agents interact with environments through sensors & actuators. A human agent has eyes, ears, & other organs for sensors & hands, legs, vocal tract, & so on for actuators. A robotic agent might have cameras & infrared range finders for sensors & various motors for actuators. A software agent receives file contents, network packets, & human input (keyboard/mouse/touchscreen/voice) as sensory inputs & acts on environment by writing files, sending network packets, & displaying information or generating sounds. Environment could be everything – entire universe! In practice it is just that part of universe whose state we care about when designing this agent – part that affects what agent perceives & is affected by agent's actions.

– *Tác nhân & Môi trường.* Một tác nhân là bất kỳ thứ gì có thể được xem như nhận thức môi trường của nó thông qua các cảm biến & tác động lên môi trường đó thông qua các bộ truyền động. Ý tưởng đơn giản này được minh họa trong Hình 2.1: Các tác nhân tương tác với môi trường thông qua các cảm biến & bộ truyền động. Một tác nhân con người có mắt, tai, & các cơ quan khác để cảm biến & tay, chân, đường thanh quản, & v.v. để truyền động. Một tác nhân rô bốt có thể có camera & máy đo khoảng cách hồng ngoại để cảm biến & nhiều động cơ khác nhau để truyền động. Một tác nhân phần mềm nhận nội dung tệp, các gói mạng, & đầu vào của con người (bàn phím/chuột/màn hình cảm ứng/giọng nói) làm đầu vào cảm biến & tác động lên môi trường bằng cách ghi tệp, gửi các gói mạng, & hiển thị thông tin hoặc tạo ra âm thanh. Môi trường có thể là tất cả mọi thứ – toàn bộ vũ trụ! Trên thực tế, chỉ có một phần của vũ trụ mà chúng ta quan tâm đến trạng thái của nó khi thiết kế tác nhân này – phần ảnh hưởng đến những gì tác nhân nhận thức & bị ảnh hưởng bởi các hành động của tác nhân.

Use term *percept* to refer to content an agent's sensors are perceiving. An agent's *percept sequence* is complete history of everything agent has ever perceived. In general, *an agent's choice of action at any given instant can depend on its built-in knowledge & on entire percept sequence observed to date, but not on anything it hasn't perceived.* By specifying agent's choice of action for every possible percept sequence, have said more or less everything there is to say about agent. Mathematically speaking, say: an agent's behavior is described by *agent function* that maps any given percept sequence to an action.

– Sử dụng thuật ngữ percept để chỉ nội dung mà các cảm biến của tác nhân đang nhận thức. Chuỗi nhận thức của tác nhân là lịch sử hoàn chỉnh về mọi thứ mà tác nhân từng nhận thức. Nhìn chung, *lựa chọn hành động của tác nhân tại*

bất kỳ thời điểm nào có thể phụ thuộc vào kiến thức tích hợp của tác nhân & vào toàn bộ chuỗi nhận thức được quan sát cho đến nay, nhưng không phụ thuộc vào bất kỳ điều gì mà tác nhân chưa nhận thức. Bằng cách chỉ định lựa chọn hành động của tác nhân cho mọi chuỗi nhận thức có thể, đã nói ít nhiều mọi thứ cần nói về tác nhân. Về mặt toán học, nói: hành vi của tác nhân được mô tả bởi *hàm tác nhân* ánh xạ bất kỳ chuỗi nhận thức nào cho một hành động.

Can imagine *tabulating* agent function that describes any given agent; for most agents, this would be a very large table – infinite, in fact, unless we place a bound on length of percept sequences we want to consider. Given an agent to experiment with, we can, in principle, construct this table by trying out all possible percept sequences & recording which actions agent does in response. [If agent uses some randomization to choose its actions, then would have to try each sequence many times to identify probability of each action. One might imagine: acting randomly is rather silly, but show later in this chap: it can be very intelligent.] Table is, of course, an *external* characterization of agent. *Internally*, agent function for an artificial agent will be implemented by an *agent program*. Important to keep these 2 ideas distinct. Agent function is an abstract mathematical description; agent program is a concrete implementation, running within some physical system.

– Có thể tưởng tượng việc lập bảng hàm tác nhân mô tả bất kỳ tác nhân nào; đối với hầu hết các tác nhân, đây sẽ là một bảng rất lớn – vô hạn, trên thực tế, trừ khi chúng ta đặt một giới hạn về độ dài của các chuỗi nhận thức mà chúng ta muốn xem xét. Với một tác nhân để thử nghiệm, về nguyên tắc, chúng ta có thể xây dựng bảng này bằng cách thử tất cả các chuỗi nhận thức có thể & ghi lại hành động mà tác nhân thực hiện để phản hồi. [Nếu tác nhân sử dụng một số ngẫu nhiên để chọn hành động của mình, thì sẽ phải thử từng chuỗi nhiều lần để xác định xác suất của từng hành động. Người ta có thể tưởng tượng: hành động ngẫu nhiên khá ngớ ngẩn, nhưng sẽ hiển thị sau trong chương này: nó có thể rất thông minh.] Tất nhiên, bảng là một đặc điểm *bên ngoài* của tác nhân. *Bên trong*, hàm tác nhân cho một tác nhân nhân tạo sẽ được triển khai bởi một *chương trình tác nhân*. Điều quan trọng là phải giữ cho 2 ý tưởng này riêng biệt. Hàm tác nhân là một mô tả toán học trừu tượng; chương trình tác nhân là một triển khai cụ thể, chạy trong một số hệ thống vật lý.

To illustrate these ideas, use a simple example – vacuum-cleaner world, which consists of a robotic vacuum-cleaning agent in a world consisting squares that can be either dirty or clean. Fig. 2.2: A vacuum-cleaner world with just 2 locations. Each location can be clean or dirty, & agent can move left or right & can clean square that it occupies. Different versions of vacuum world allow for different rules about what agent can perceive, whether its actions always succeed, & so on. shows a configuration with just 2 squares, A & B. Vacuum agent perceives which square it is in & whether there is dirt in square. Agent starts in square A. Available actions are to move to right, move to left, suck up dirt, or do nothing. [In a real robot, it would be unlikely to have an actions like "move right" & "move left". Instead actions would be "spin wheels forward" & "spin wheels backward". Have chosen actions to be easier to follow on page, not for ease of implementation in an actual robot.] 1 very simple agent function is following: if current square is dirty, then suck; otherwise, move to other square. A partial tabulation of this agent function is shown in Fig. 2.3: Partial tabulation of a simple agent function for vacuum-cleaner world shown in Fig. 2.2. Agent cleans current square if it is dirty, otherwise it moves to other square. Note: table is of unbounded size unless there is a restriction on length of possible percept sequences. & an agent program that implements it appears in Fig. 2.8: Agent program for a simple reflex agent in 2-location vacuum environment. This program implements agent function tabulated in Fig. 2.3.

Looking at Fig. 2.3, see: various vacuum-world agents can be defined simply by filling in RH column in various ways. Obvious question, then: *What is right way to fill out table?* I.e., what makes an agent good or bad, intelligent or stupid? Answer these questions in next sect.

Before close this sect, should emphasize: notion of an agent is meant to be a tool for analyzing systems, not an absolute characterization that divides world into agents & non-agents. One could view a hand-held calculator as an agent that chooses action of displaying "4" when given percept sequence "$2 + 2 =$", but such an analysis would hardly aid our understanding of calculator. In a sense, all areas of engineering can be seen as designing artifacts that interact with world; AI operates at (what authors consider to be) most interesting end of spectrum, where artifacts have significant computational resources & task environment requires nontrivial decision making.

– Trước khi kết thúc phần này, cần nhấn mạnh: khái niệm về tác nhân được hiểu là một công cụ để phân tích hệ thống, không phải là một đặc điểm tuyệt đối chia thế giới thành tác nhân & không phải tác nhân. Người ta có thể xem máy tính cầm tay như một tác nhân chọn hành động hiển thị "4" khi đưa ra chuỗi nhận thức "$2 + 2 =$", nhưng một phân tích như vậy khó có thể giúp chúng ta hiểu máy tính. Theo một nghĩa nào đó, tất cả các lĩnh vực kỹ thuật đều có thể được coi là thiết kế các hiện vật tương tác với thế giới; AI hoạt động ở (những gì các tác giả coi là) phần cuối của quang phổ thú vị nhất, nơi các hiện vật có tài nguyên tính toán đáng kể & môi trường tác vụ đòi hỏi phải đưa ra quyết định không tầm thường.

∗ 2.2. Good Behavior: Concept of Rationality. A *rational agent* is one that does right thing. Obviously, doing right thing is better than doing wrong thing, but what does it mean to do right thing?

· 2.2.1. Performance measures. Moral philosophy has developed several different notions of "right thing", but AI has generally stuck to 1 notion called *consequentialism*: evaluate an agent's behavior by its consequences. When an agent is plunked down in an environment, it generates a sequence of actions according to percepts it receives. This sequence of actions causes environment to go through a sequence of states. If sequence is desirable, then agent has performed well. This notion of desirability is captured by a *performance measure* that evaluates any given sequence of environment states.

– *Các biện pháp hiệu suất.* Triết học đạo đức đã phát triển một số khái niệm khác nhau về "điều đúng đắn", nhưng AI thường gắn bó với 1 khái niệm gọi là chủ nghĩa hậu quả: đánh giá hành vi của tác nhân theo hậu quả của nó. Khi một tác nhân được đặt xuống trong một môi trường, nó sẽ tạo ra một chuỗi hành động theo các nhận thức mà nó nhận được. Chuỗi hành động này khiến môi trường trải qua một chuỗi trạng thái. Nếu chuỗi là mong muốn, thì tác nhân đã

hoạt động tốt. Khái niệm mong muốn này được nắm bắt bằng một *biện pháp hiệu suất* đánh giá bất kỳ chuỗi trạng thái môi trường nào được đưa ra.

Humans have desires & preferences of their own, so notion of rationality as applied to humans has to do with their success in choosing actions that produce sequences of environment states that are desirable *from their point of view.* Machines, on other hand, do not have desires & preferences of their own; performance measure is, initially at least, in mind of designer of machine, or in mind of users machine is designed for. See: some agent designs have an explicit representation of (a version of) performance measure, while in other designs performance measure is entirely implicit – agent may do right thing, but it doesn't know why.

– Con người có ham muốn & sở thích riêng, vì vậy khái niệm về tính hợp lý khi áp dụng cho con người có liên quan đến thành công của họ trong việc lựa chọn các hành động tạo ra chuỗi trạng thái môi trường mong muốn *theo quan điểm của họ.* Mặt khác, máy móc không có ham muốn & sở thích riêng; thước đo hiệu suất, ít nhất là ban đầu, nằm trong tâm trí của nhà thiết kế máy móc hoặc trong tâm trí của người dùng mà máy móc được thiết kế cho. Xem: một số thiết kế tác nhân có biểu diễn rõ ràng về (một phiên bản) thước đo hiệu suất, trong khi trong các thiết kế khác, thước đo hiệu suất hoàn toàn ngầm định – tác nhân có thể làm đúng, nhưng không biết tại sao.

Recalling Norbert Wiener's warning to ensure "purpose put into machine is purpose which we really desire", notice: it can be quite hard to formulate a performance measure correctly. Consider, e.g., vacuum–cleaner agent from preceding sect. Might propose to measure performance by amount of dirt cleaned up in a single 8-hour shift. With a rational agent, of course, what you ask for is what you get. A rational agent can maximize this performance measure by cleaning up dirt, then dumping it all on floor, then cleaning it up again, & so on. A more suitable performance measure would reward agent for having a clean floor. E.g., 1 point could be awarded for each clean square at each time step (perhaps with a penalty for electricity consumed & noise generated). *As a general rule, better to design performance measures according to what one actually wants to be achieved in environment, rather than according to how one thinks agent should behave.*

– Nhắc lại lời cảnh báo của Norbert Wiener để đảm bảo "mục đích đưa vào máy là mục đích mà chúng ta thực sự mong muốn", hãy lưu ý: có thể khá khó để xây dựng một thước đo hiệu suất chính xác. Ví dụ, hãy xem xét tác nhân máy hút bụi từ phần trước. Có thể đề xuất đo hiệu suất theo lượng bụi bẩn được làm sạch trong một ca làm việc 8 giờ. Tất nhiên, với một tác nhân hợp lý, những gì bạn yêu cầu là những gì bạn nhận được. Một tác nhân hợp lý có thể tối đa hóa thước đo hiệu suất này bằng cách làm sạch bụi bẩn, sau đó đổ hết xuống sàn, rồi lại làm sạch, & cứ như vậy. Một thước đo hiệu suất phù hợp hơn sẽ thưởng cho tác nhân vì có sàn sạch. Ví dụ, có thể thưởng 1 điểm cho mỗi ô vuông sạch tại mỗi bước thời gian (có thể kèm theo hình phạt cho lượng điện tiêu thụ & tiếng ồn tạo ra). *Theo nguyên tắc chung, tốt hơn là thiết kế các thước đo hiệu suất theo những gì người ta thực sự muốn đạt được trong môi trường, thay vì theo cách người ta nghĩ tác nhân nên hành xử.*

Even when obvious pitfalls are avoided, some knotty problems remain. E.g., notion of "clean floor" in preceding paragraph is based on average cleanliness over time. Yet same average cleanliness can be achieved by 2 different agents, one of which does a mediocre job all time while other cleans energetically but takes long breaks. Which is preferable might seem to be a fine point of janitorial science, but in fact it is a deep philosophical question with far-reaching implications. Which is better – an economy where everyone lives in moderate poverty, or one in which some live in plenty while others are very poor? Leave these questions as an exercise for diligent reader.

– Ngay cả khi tránh được những cạm bẫy rõ ràng, một số vấn đề nan giải vẫn còn tồn tại. Ví dụ, khái niệm "sàn nhà sạch" trong đoạn trước dựa trên mức độ sạch trung bình theo thời gian. Tuy nhiên, mức độ sạch trung bình tương tự có thể đạt được bởi 2 tác nhân khác nhau, một trong số đó làm việc tầm thường mọi lúc trong khi tác nhân kia làm việc rất hăng hái nhưng lại nghỉ giải lao rất lâu. Cái nào tốt hơn có vẻ là một điểm tinh tế của khoa học vệ sinh, nhưng trên thực tế, đó là một câu hỏi triết học sâu sắc với những hàm ý sâu xa. Cái nào tốt hơn – một nền kinh tế mà mọi người đều sống trong cảnh nghèo đói vừa phải, hay một nền kinh tế mà một số người sống trong cảnh sung túc trong khi những người khác lại rất nghèo? Hãy để những câu hỏi này như một bài tập cho người đọc siêng năng.

For most of book, assume: performance measure can be specified correctly. For reasons given above, however, must accept possibility that we might put wrong purpose into machine – precisely King Midas problem described on p. 51. Moreover, when designing 1 piece of software, copies of which will belong to different users, cannot anticipate exact preferences of each individual user. Thus, may need to build agents that reflect initial uncertainty about true performance measure & learn more about it as time goes by; such agents are described in Chaps. 15, 17, 23.

· 2.2.2. Rationality. What is rational at any given time depends on 4 things:

1. Performance measure that defines criterion of success.
2. Agent's prior knowledge of environment.
3. Actions that agent can perform.
4. Agent's percept sequence to date.

This leads to a def of a rational agent:

Definition 1 (Rational agent)**.** *For each possible percept sequence, a* rational agent *should select an action that is expected to maximize its performance measure, given evidence provided by percept sequence & whatever built-in knowledge agent has.*

– Đối với mỗi chuỗi nhận thức có thể, một tác nhân hợp lý *nên chọn một hành động dự kiến sẽ tối đa hóa thước đo hiệu suất của nó, dựa trên bằng chứng được cung cấp bởi chuỗi nhận thức & bất kỳ tác nhân kiến thức tích hợp nào.*

Consider simple vacuum-cleaner agent that cleans a square if it is dirty & moves to other square if not; this is agent function tabulated in Fig. 2.3. Is this a rational agent? That depends! 1st, need to say what performance measure is, what is known about environment, & what sensors & actuators agent has. Assume:

1. Performance measure awards 1 point for each clean square at each time step, over a "lifetime" of 1000 time steps.
2. "Geography" of environment is known *a priori* Fig. 2.2 but dirt distribution & initial location of agent are not. Clean squares stay clean & sucking cleans current square. *Right & Left* actions move agent 1 square except when this would take agent outside environment, in which case agent remains where it is.
3. Only available actions are *Right, Left, & Suck*.
4. Agent correctly perceives its location & whether that location contains dirt.

Under these circumstances agent is indeed rational; its expected performance is at least as good as other agent's.
One can see easily: same agent would be irrational under different circumstances. E.g., once all dirt is cleaned up, agent will oscillate needlessly back & forth; if performance measure includes a penalty of 1 point for each movement, agent will fare poorly.tA better agent for this case would do nothing once it is sure: all squares are clean. If clean squares can become dirty again, agent should occasionally check & reclean them if needed. If geography of environment is unknown, agent will need to *explore* it. Exercise 2.VACR asks you to design agents for these cases.
– Người ta có thể dễ dàng thấy: cùng một tác nhân sẽ không hợp lý trong những hoàn cảnh khác nhau. Ví dụ, sau khi tất cả bụi bẩn được dọn sạch, tác nhân sẽ dao động qua lại không cần thiết &; nếu thước đo hiệu suất bao gồm hình phạt 1 điểm cho mỗi chuyển động, tác nhân sẽ hoạt động kém. Một tác nhân tốt hơn cho trường hợp này sẽ không làm gì cả khi chắc chắn: tất cả các ô vuông đều sạch. Nếu các ô vuông sạch có thể bị bẩn trở lại, tác nhân nên thỉnh thoảng kiểm tra & làm sạch lại chúng nếu cần. Nếu địa lý của môi trường không xác định, tác nhân sẽ cần phải *khám phá* nó. Bài tập 2.VACR yêu cầu bạn thiết kế các tác nhân cho những trường hợp này.

· 2.2.3. Omniscience, learning, & autonomy. Need to be careful to distinguish between rationality & *omniscience*. An omniscient agent knows *actual* outcome of its actions & can act accordingly; but omniscience is impossible in reality. Consider example: I am walking along Champs Elysées 1 day & I see an old friend across street. There is no traffic nearby & I'm not otherwise engaged, so, being rational, I start to cross street. Meanwhile, at 33000 feet, a cargo door falls off a passing airliner [See N. HENDERSON, "New door latches urged for Boeing 747 jumbo jets," Washington Post, Aug 24, 1989.], & before I make it to other side of street I am flattened. Was I irrational to cross street? It is unlikely: my obituary would read "Idiot attempts to cross street."
This example shows: rationality is not same as perfection. Rationality maximizes *expected* performance, while perfection maximizes *actual* performance. Retreating from a requirement of perfection is not just a question of being fair to agents. Point: if expect an agent to do what turns out after fact to be best action, it will be impossible to design an agent to fulfill this specification – unless improve performance of crystal balls or time machines.
– Ví dụ này cho thấy: tính hợp lý không giống với sự hoàn hảo. Tính hợp lý tối đa hóa hiệu suất *mong đợi*, trong khi sự hoàn hảo tối đa hóa hiệu suất *thực tế*. Việc rút lui khỏi yêu cầu về sự hoàn hảo không chỉ là vấn đề công bằng với các tác nhân. Điểm chính: nếu mong đợi một tác nhân thực hiện những gì sau này trở thành hành động tốt nhất, thì sẽ không thể thiết kế một tác nhân để đáp ứng thông số kỹ thuật này – trừ khi cải thiện hiệu suất của quả cầu pha lê hoặc cỗ máy thời gian.
Our definition of rationality does not require omniscience, then, because rational choice depends only on percept sequence *to date*. Must also ensure: we haven't inadvertently allowed agent to engage in decidedly underintelligent activities. E.g., if an agent does not look both ways before crossing a busy road, then its percept sequence will not tell it that there is a large truck approaching at high speed. Does our definition of rationality say that it's now OK to cross road? Far from it!
– Định nghĩa của chúng ta về tính hợp lý không đòi hỏi sự toàn năng, vì sự lựa chọn hợp lý chỉ phụ thuộc vào chuỗi nhận thức *cho đến nay*. Cũng phải đảm bảo: chúng ta không vô tình cho phép tác nhân tham gia vào các hoạt động rõ ràng là thiếu thông minh. Ví dụ, nếu một tác nhân không nhìn cả hai hướng trước khi băng qua một con đường đông đúc, thì chuỗi nhận thức của tác nhân đó sẽ không cho tác nhân biết rằng có một chiếc xe tải lớn đang lao tới với tốc độ cao. Định nghĩa của chúng ta về tính hợp lý có nói rằng bây giờ có thể băng qua đường không? Hoàn toàn không phải vậy!
1st, it would not be rational to cross road given this uninformative percept sequence: risk of accident from crossing without looking is too great. 2nd, a rational agent should choose "looking" action before stepping into street, because looking helps maximize expected performance. Doing actions *in order to modify future percepts* – sometimes called *information gathering* – is an important part of rationality & is covered in depth in Chap. 15. A 2nd example of information gathering is provided by *exploration* that must be undertaken by a vacuum-cleaning agent in an initially unknown environment.
– 1. Sẽ không hợp lý khi băng qua đường khi xét đến chuỗi nhận thức không cung cấp thông tin này: nguy cơ tai nạn khi băng qua đường mà không nhìn là quá lớn. 2. Một tác nhân hợp lý nên chọn hành động "nhìn" trước khi bước vào đường, vì nhìn giúp tối đa hóa hiệu suất mong đợi. Thực hiện các hành động *để sửa đổi các nhận thức trong tương lai* – đôi khi được gọi là *thu thập thông tin* – là một phần quan trọng của tính hợp lý & được trình bày sâu trong Chương 15. Ví dụ thứ 2 về việc thu thập thông tin được cung cấp bởi *khám phá* phải được thực hiện bởi một tác nhân hút bụi trong một môi trường ban đầu không xác định.
Our definition requires a rational agent not only to gather information but also to *learn* as much as possible from what it perceives. Agent's initial configuration could reflect some prior knowledge of environment, but as agent gains experience this may be modified & augmented. There are extreme cases in which environment is completely known *a priori* & completely predictable. In such cases, agent need not perceive or learn; it simply acts correctly.
– Định nghĩa của chúng tôi yêu cầu một tác nhân lý trí không chỉ thu thập thông tin mà còn *học* càng nhiều càng tốt từ những gì nó nhận thức. Cấu hình ban đầu của tác nhân có thể phản ánh một số kiến thức trước đó về môi trường,

nhưng khi tác nhân có thêm kinh nghiệm, điều này có thể được sửa đổi & tăng cường. Có những trường hợp cực đoan trong đó môi trường được biết đến hoàn toàn *a priori* & hoàn toàn có thể dự đoán được. Trong những trường hợp như vậy, tác nhân không cần phải nhận thức hoặc học; nó chỉ đơn giản là hành động đúng.

Of course, such agents are fragile. Consider lowly dung beetle. After digging its nest & laying its eggs, it fetches a ball of dung from a nearby heap to plug entrance. If ball of dung is removed from its grasp *en route*, beetle continues its task & pantomimes plugging nest with nonexistent dung ball, never noticing that it is missing. Evolution has built an assumption into beetle's behavior, & when it is violated, unsuccessful behavior results.

– Tất nhiên, những tác nhân như vậy rất mong manh. Hãy xem xét loài bọ hung thấp kém. Sau khi đào tổ & đẻ trứng, nó lấy một cục phân từ đống phân gần đó để chặn lối vào. Nếu cục phân bị lấy khỏi tay nó trên đường đi, bọ hung tiếp tục nhiệm vụ & làm trò hề bằng cách chặn tổ bằng cục phân không tồn tại, không bao giờ nhận ra rằng cục phân đã mất. Sự tiến hóa đã xây dựng một giả định vào hành vi của bọ hung, & khi giả định đó bị vi phạm, hành vi không thành công sẽ xảy ra.

Slightly more intelligent is sphex wasp. Female sphex will dig a burrow, go out & sting a caterpillar & drag it to burrow, enter burrow again to check all is well, drag caterpillar inside, & lay its eggs. Caterpillar serves as a food source when eggs hatch. So far so good, but if an entomologist moves caterpillar a few inches away while sphex is doing check, it will revert to "drag caterpillar" step of its plan & will continue plan without modification, re–checking burrow, even after dozens of caterpillar-moving interventions. Sphex is unable to learn that its innate plan is failing, & thus will not change it.

– Thông minh hơn một chút là ong bắp cày sphex. Ong sphex cái sẽ đào hang, chui ra & đốt sâu bướm & kéo nó vào hang, chui vào hang lần nữa để kiểm tra mọi thứ ổn thỏa, kéo sâu bướm vào trong, & đẻ trứng. Sâu bướm đóng vai trò là nguồn thức ăn khi trứng nở. Cho đến giờ thì mọi thứ vẫn ổn, nhưng nếu một nhà côn trùng học di chuyển sâu bướm ra xa vài inch trong khi sphex đang kiểm tra, nó sẽ quay lại bước "kéo sâu bướm" trong kế hoạch của nó & sẽ tiếp tục kế hoạch mà không cần sửa đổi, kiểm tra lại hang, ngay cả sau hàng chục lần can thiệp di chuyển sâu bướm. Sphex không thể học được rằng kế hoạch bẩm sinh của nó đang thất bại, & do đó sẽ không thay đổi nó.

To extent that an agent relies on prior knowledge of its designer rather than on its own percepts & learning processes, say: agent lacks *autonomy*. A rational agent should be autonomous – it should learn what it can to compensate for partial or incorrect prior knowledge. E.g., a vacuum-cleaning agent that learns to predict where & when additional dirt will appear will do better than one that does not.

– Trong phạm vi mà một tác nhân dựa vào kiến thức trước đó của người thiết kế nó hơn là vào các nhận thức & quá trình học tập của chính nó, hãy nói: tác nhân thiếu *tự chủ*. Một tác nhân hợp lý phải tự chủ – nó phải học những gì nó có thể để bù đắp cho kiến thức trước đó không đầy đủ hoặc không chính xác. Ví dụ, một tác nhân hút bụi học cách dự đoán nơi & khi nào bụi bẩn sẽ xuất hiện sẽ hoạt động tốt hơn một tác nhân không làm như vậy.

As a practical matter, one seldom requires complete autonomy from start: when agent has had little or no experience, it would have to act randomly unless designer gave some assistance. Just as evolution provides animals with enough built-in reflexes to survive long enough to learn for themselves, it would be reasonable to provide an AI agent with some initial knowledge as well as an ability to learn. After sufficient experience of its environment, behavior of a rational agent can become effectively *independent* of its prior knowledge. Hence, incorporation of learning allows one to design a single rational agent that will succeed in a vast variety of environments.

– Trên thực tế, người ta hiếm khi đòi hỏi sự tự chủ hoàn toàn ngay từ đầu: khi tác nhân có ít hoặc không có kinh nghiệm, nó sẽ phải hành động ngẫu nhiên trừ khi nhà thiết kế cung cấp một số hỗ trợ. Cũng giống như quá trình tiến hóa cung cấp cho động vật đủ phản xạ tích hợp để tồn tại đủ lâu để tự học, sẽ hợp lý khi cung cấp cho tác nhân AI một số kiến thức ban đầu cũng như khả năng học hỏi. Sau khi có đủ kinh nghiệm về môi trường của mình, hành vi của tác nhân hợp lý có thể trở nên *độc lập* hiệu quả với kiến thức trước đó của nó. Do đó, việc kết hợp học tập cho phép người ta thiết kế một tác nhân hợp lý duy nhất sẽ thành công trong nhiều môi trường khác nhau.

∗ 2.3. Nature of Environments. Now have a definition of rationality, almost ready to think about building rational agents. 1st, however, must think about *task environments*, which are essentially "problems" to which rational agents are "solutions". Begin by showing how to specify a task environment, illustrating process with a number of examples. Then show: task environments come in a variety of flavors. Nature of task environment directly affects appropriate design for agent program.

– *Bản chất của Môi trường.* Bây giờ đã có định nghĩa về tính hợp lý, gần như đã sẵn sàng để nghĩ về việc xây dựng các tác nhân hợp lý. Tuy nhiên, trước tiên, phải nghĩ về *task environments*, về cơ bản là "các vấn đề" mà các tác nhân hợp lý là "giải pháp". Bắt đầu bằng cách chỉ ra cách chỉ định một môi trường tác vụ, minh họa quy trình bằng một số ví dụ. Sau đó chỉ ra: các môi trường tác vụ có nhiều loại khác nhau. Bản chất của môi trường tác vụ ảnh hưởng trực tiếp đến thiết kế phù hợp cho chương trình tác nhân.

· 2.3.1. Specifying task environment. In our discussion of rationality of simple vacuum-cleaner agent, had to specify performance measure, environment, & agent's actuators & sensors. Group all these under heading of *task environment*. For acronymically minded, call this PEAS (Performance, Environment, Actuators, Sensors) description. In designing an agent, 1st step mus always be to specify task environment as fully as possible.

Vacuum world was a simple example; consider a more complex problem: an automated taxi driver. Fig. 2.4: PEAS description of task environment for an automated taxi driver. summarizes PEAS description for taxi's task environment. Discuss each element in more detail.

1st, what is *performance measure* to which we would like our automated driver to aspire? Desirable qualities include getting to correct destination; minimizing fuel consumption & wear & tear; minimizing trip time or cost; minimizing violations of traffic laws & disturbances to other drivers; maximizing safety & passenger comfort; maximizing profits.

Obviously, some of these goals conflict, so tradeoffs will be required.
– 1, *tiêu chuẩn hiệu suất* mà chúng ta muốn trình điều khiển tự động của mình hướng tới là gì? Các phẩm chất mong muốn bao gồm đến đúng đích; giảm thiểu mức tiêu thụ nhiên liệu & hao mòn & rách; giảm thiểu thời gian hoặc chi phí chuyến đi; giảm thiểu vi phạm luật giao thông & gây phiền nhiễu cho những người lái xe khác; tối đa hóa sự an toàn & sự thoải mái của hành khách; tối đa hóa lợi nhuận. Rõ ràng, một số mục tiêu này xung đột với nhau, vì vậy cần phải đánh đổi.
Next, what is driving *environment* that taxi will face? Any taxi driver must deal with a variety of roads, ranging from rural lanes & urban alleys to 12-lane freeways. Roads contain other traffic, pedestrians, stray animals, road works, police cars, puddles, & potholes. Taxi must also interact with potential & actual passengers. There are also some optional choices. Taxi might need to operate in Southern California, where snow is seldom a problem, or in Alaska, where it seldom is not. It could always driving on right, or we might want it to be flexible enough to drive on left when in Britain or Japan. Obviously, more restricted environment, easier design problem.
– Tiếp theo, môi trường lái xe mà taxi sẽ phải đối mặt là gì? Bất kỳ tài xế taxi nào cũng phải đối mặt với nhiều loại đường khác nhau, từ làn đường nông thôn & ngõ phố đến đường cao tốc 12 làn. Đường có nhiều phương tiện giao thông khác, người đi bộ, động vật hoang dã, công trình đường bộ, xe cảnh sát, vũng nước, & ổ gà. Taxi cũng phải tương tác với hành khách tiềm năng & thực tế. Ngoài ra còn có một số lựa chọn tùy chọn. Taxi có thể cần hoạt động ở Nam California, nơi tuyết hiếm khi là vấn đề, hoặc ở Alaska, nơi tuyết hiếm khi không là vấn đề. Nó luôn có thể lái xe bên phải, hoặc chúng ta có thể muốn nó đủ linh hoạt để lái xe bên trái khi ở Anh hoặc Nhật Bản. Rõ ràng, môi trường hạn chế hơn, vấn đề thiết kế dễ dàng hơn.
Actuators for an automated taxi include those available to a human driver: control over engine through accelerator & control over steering & braking. In addition, it will need output to a display screen or voice synthesizer to talk back to passengers, & perhaps some way to communicate with other vehicles, politely or otherwise.
– *Bộ truyền động* cho một chiếc taxi tự động bao gồm những bộ truyền động có sẵn cho người lái: điều khiển động cơ thông qua chân ga & điều khiển tay lái & phanh. Ngoài ra, nó sẽ cần xuất ra màn hình hiển thị hoặc bộ tổng hợp giọng nói để nói chuyện với hành khách, & có lẽ là một số cách để giao tiếp với các phương tiện khác, theo cách lịch sự hoặc không.
Basic *sensors* for taxi will include 1 or more video cameras so that it can see, as well as lidar & ultrasound sensors to detect distances to other cars & obstacles. To avoid speeding tickets, taxi should have a speedometer, & to control vehicle properly, especially on curves, it should have an accelerometer. To determine mechanical state of vehicle, it will need usual array of engine, fuel, & electrical system sensors. Like many human drivers, it might want to access GPT signals so that it doesn't get lost. Finally, it will need touchscreen or voice input for passenger to request a destination.
– Các cảm biến cơ bản của taxi sẽ bao gồm 1 hoặc nhiều camera video để có thể nhìn thấy, cũng như cảm biến lidar & siêu âm để phát hiện khoảng cách đến các xe khác & chướng ngại vật. Để tránh bị phạt vì chạy quá tốc độ, taxi nên có đồng hồ đo tốc độ, & để điều khiển xe đúng cách, đặc biệt là khi vào cua, taxi nên có máy đo gia tốc. Để xác định trạng thái cơ học của xe, taxi sẽ cần một loạt các cảm biến thông thường về động cơ, nhiên liệu, & hệ thống điện. Giống như nhiều tài xế khác, taxi có thể muốn truy cập tín hiệu GPT để không bị lạc đường. Cuối cùng, taxi sẽ cần màn hình cảm ứng hoặc giọng nói để hành khách yêu cầu điểm đến.
In Fig. 2.5: Examples of agent types & their PEAS descriptions., have sketched basic PEAS elements for a number of additional agent types. Further examples appear in Exercise 2.PEAS. Examples include physical as well as virtual environments. Note: virtual task environments can be just as complex as "real" world: e.g., a *software agent* (or software robot or *softbot*) that trades on auction & reselling Web sites deals with millions of other users & billions of objects, many with real images.
– Trong Hình 2.5: Ví dụ về các loại tác nhân & mô tả PEAS của chúng., đã phác thảo các thành phần PEAS cơ bản cho một số loại tác nhân bổ sung. Các ví dụ khác xuất hiện trong Bài tập 2.PEAS. Các ví dụ bao gồm cả môi trường vật lý cũng như môi trường ảo. Lưu ý: môi trường tác vụ ảo có thể phức tạp như thế giới "thực": ví dụ, một *phần mềm tác nhân* (hoặc rô-bốt phần mềm hoặc *softbot*) giao dịch trên các trang web đấu giá & bán lại giao dịch với hàng triệu người dùng khác & hàng tỷ đối tượng, nhiều đối tượng có hình ảnh thực.

· 2.3.2. Properties of task environments. Range of task environments that might arise in AI is obviously vast. Can, however, identify a fairly small number of dimensions along which task environments can be categorized. These dimensions determine, to a large extent, appropriate agent design & applicability of each of principal families of techniques for agent implementation. 1st list dimensions, then analyze several task environments to illustrate ideas. Definitions here are informal; later chaps provide more precise statements & examples of each kind of environment.
– *Thuộc tính của môi trường tác vụ.* Phạm vi các môi trường tác vụ có thể phát sinh trong AI rõ ràng là rất lớn. Tuy nhiên, có thể xác định một số lượng khá nhỏ các chiều mà môi trường tác vụ có thể được phân loại theo. Các chiều này xác định, ở một mức độ lớn, thiết kế tác nhân phù hợp & khả năng áp dụng của từng họ kỹ thuật chính để triển khai tác nhân. Đầu tiên, hãy liệt kê các chiều, sau đó phân tích một số môi trường tác vụ để minh họa các ý tưởng. Các định nghĩa ở đây là không chính thức; các chương sau cung cấp các tuyên bố chính xác hơn & ví dụ về từng loại môi trường. p. 61+++

◦

PART II: PROBLEM-SOLVING.

◦ 3. Solving Problems by Searching.

◦ 4. Search in Complex Environments.

- 5. Constraint Satisfaction Problems.
- 6. Adversarial Search & Games.

PART III: KNOWLEDGE, REASONING, & PLANNING.

- 7. Logical Agents.
- 8. 1st-Order Logic.
- 9. Inference in 1st-Order Logic.
- 10. Knowledge Representation.
- 11. Automated Planning.

PART IV: UNCERTAIN KNOWLEDGE & REASONING.

- 12. Quantifying Uncertainty.
- 13. Probabilistic Reasoning.
- 14. Probabilistic Reasoning over Time.
- 15. Making Simple Decisions.
- 16. Making Complex Decisions.
- 17. Multiagent Decision Making.
- 18. Probabilistic Programming.

PART V: ML

- 19. Learning from Examples.
- 20. Knowledge in Learning.
- 21. Learning Probabilistic Models.
- 22. DL.
- 23. Reinforcement Learning.

PART VI: COMMUNICATING, PERCEIVING, & ACTING.

- 24. Natural Language Processing.
- 25. DL for NLP.
- 26. Robotics.
- 27. Computer Vision.

PART VII: CONCLUSIONS.

- 28. Philosophy, Ethics, & Safety of AI.
- 29. Future of AI.

- Appendix A: Mathematical Background.
- Appendix B: Notes on Languages & Algorithms.

2 Miscellaneous

Tài liệu

[NR21] Peter Norvig and Stuart Russell. *Artificial Intelligence: A Modern Approach*. 4th Edition, Global Edition. Pearson Series In Artificial Intelligence. Pearson, 2021, p. 1166.