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: .pdf.
  TEX: URL: .tex.
- .
  PDF: URL: .pdf.
  TEX: URL: .tex.

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.
    * 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.

*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.

* 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.

  * 1.2. Foundations of AI.
- 2. Intelligent Agents.

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.