第一讲 人工智能的哲学之门

潘斌　　华东师范大学哲学系教授

一 人工智能的前世今生

1. 中国古代人工智能实验

周穆王西巡狩，越昆仑，不至弇山。反还，未及中国，道有献工人名偃师。穆王荐之，问曰：“若有何能？”偃师曰：“臣唯命所试。然臣已有所造，愿王先观之。”穆王曰：“日以俱来，吾与若俱观之。”翌日偃师谒见王。王荐之，曰：“若与偕来者何人邪？”对曰：“臣之所造能倡者。”穆王惊视之，趋步俯仰，信人也。巧夫！领其颅，则歌合律；捧其手，则舞应节。千变万化，惟意所适。王以为实人也，与盛姬内御并观之。技将终，倡者瞬其目而招王之左右侍妾。王大怒，立欲诛偃师。偃师大慑，立剖散倡者以示王，皆傅会革、木、胶、漆、白、黑、丹、青之所为。王谛料之，内则肝胆、心肺、脾肾、肠胃，外则筋骨、支节、皮毛、齿发，皆假物也，而无不毕具者。合会复如初见。王试废其心，则口不能言；废其肝，则目不能视；废其肾，则足不能步。穆王始悦而叹曰：“人之巧乃可与造化者同功乎？”诏贰车载之以归。

1. 中国古代人工智能实验

夫班输之云梯，墨翟之飞鸢，自谓能之极也。弟子东门贾、禽滑釐闻偃师之巧以告二子，二子终身不敢语艺，而时执规矩。”《列子·汤问》

人工智能在中国的历史渊源:

- ⑩司辰、击鼓、 报时的“机关人”，
- ⑩会跳舞的“人形舞姬”，
- ⑩能捕鼠的木制“钟馗”，
- ⑩会化缘的“木僧人”，
- ⑩诸葛亮制作的“木牛流马”

AI诞生：达特茅斯会议

- AI诞生于一次历史性的聚会
- 时间：1956年夏季
- 地点：达特莫斯（Dartmouth）大学
- 目的：为使计算机变得更“聪明”，或者说使计算机具有智能
- 发起人：
- 麦卡锡(McCarthy），Dartmouth的年轻数学家、计算机专家，后为MIT教授
- 明斯基(M.L.Minsky），哈佛大学数学家、神经学家，后为MIT教授
- 洛切斯特(N.Lochester)， IBM公司信息中心负责人
- 香农(C.E.Shannon)，贝尔实验室信息部数学研究员
- 参加人：
- 莫尔(T.more)、塞缪尔(A.L.Samuel)， IBM公司
- 塞尔夫里奇(O.Selfridge)、索罗蒙夫(R.Solomonff)， MIT
- 纽厄尔(A.Newell)，兰德(RAND)公司
- 西蒙(H.A.Simon)，卡内基(Carnagie)工科大学
- 会议结果：经由麦卡锡提议正式采用了“Artificial Intelligence”这一术语

人工智能的多元视角

马文·明斯基(Marvin Minsky):

“人工智能就是让机器来完成那些如果由人来做则需要智能的事情的科学”

- “人工智能之父”
- 1927～
- 1969年获图灵奖,
- 获此殊荣的第一位人工智能学者
- 1969年获图灵奖, 1991年获IJCAI终身成就奖。他在人工智能、认知心理学、数学、计算语言学、机器人学等领域都做出了杰出贡献。他创建了MIT的AI实验室、还是MIT的Media实验室奠基人。

人工智能的历史渊源

01 1956年前

公元前384-322 亚里士多德（Aristotle）形式逻辑 三段论

02 20世纪30~40年代

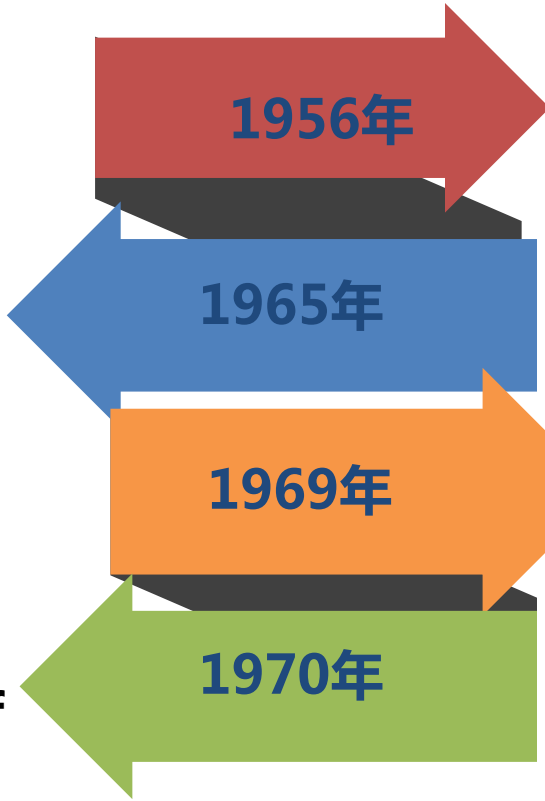
数理逻辑、维纳弗雷治、罗素等为代表对发展数理逻辑学科的贡献
丘奇(Church)、图灵和其它一些人关于计算本质的思想，为人工智能的形成产生了重要影响

03 1943年

麦卡洛克和皮茨 神经网络模型→连接主义学派

04 1948年

维纳 控制论→行为主义学派



人类历史上第一次人工智能研讨会在美国的达特茅斯大学举行，标志着人工智能学科的诞生。

Feigenbaum
专家系统 DENDRAL

召开了第一届国际人工智能联合会议，此后每两年召开一次。

《人工智能》国际杂志(International Journal of AI)创刊。

- 1956年夏出席达特茅斯会议的部分代表于50年后重逢

2006，AI 50周年会议（美国）
莫尔，麦卡锡，明斯基，塞尔夫里奇，索罗蒙夫

2

人工智能研究方法的论争及其反思

Cognition modeling (认知建模)
Knowledge Representation(知识表示)
Knowledge Reasoning(知识推理)

Knowledge Application(知识应用)
Machine Perception(机器感知)
Machine thinking(机器思维)

Machine learning(机器学习)
Machine behavior(机器行为)
Intelligent system constructing(智能系统构建) I

人工智能研究方法

* **符号主义**：以符号处理为核心的思维计算。
* **联接主义**：神经网络及神经网络间的连接机制与学习算法
* **行为主义**：智能行为的基础是“感知-行动”，是在与环境的交互作用中表现出来的。

人工智能研究方法之一：符号主义

符号主义，又称逻辑主义、心理学派或计算机学派，其学院派代表人物是纽厄尔、西蒙和尼尔逊等，强调以符号处理为核心的方法，又称为自上而下和符号主义,起源于GPS,用于模拟人类问题求解过程的心理过程,逐渐形成为物理符号系统。

AI的目标就是实现机器智能,而计算机自身具有符号处理功能,它本身就蕴含着推理能力,因而可能够方便地模拟逻辑思维过程。符号主义认为：人类智能的基本单元是符号，认知过程就是符号操作过程，从而思维就是符号计算。

符号主义的发展历程

（一）奠基与先驱 (1950s - 1960s)

•核心思想: 智能 = 符号处理 + 逻辑推理

•代表人物及其贡献 (一一对应):

- **人物:** 艾伦·纽厄尔 & 赫伯特·西蒙 (Allen Newell & Herbert Simon)
  - **机构:** 卡内基梅隆大学 (CMU) / 兰德公司 (RAND Corp.)
  - **技术/理论:** 逻辑理论家 (Logic Theorist - 首个AI程序), GPS (通用问题求解器 / 手段-目的分析 / Heuristic Search应用)
- **人物:** 约翰·麦卡锡 (John McCarthy)
  - **机构:** 达特茅斯 / MIT / 斯坦福 (Dartmouth / MIT / Stanford)
  - **技术/理论:** LISP 语言发明 (符号处理核心语言), "人工智能" 术语提出 (AI领域命名)
- **人物:** 马文·明斯基 (Marvin Minsky)
  - **机构:** MIT 人工智能实验室 (MIT AI Lab)
  - **技术/理论:** AI早期理论构建与倡导 (Foundational AI theory & advocacy), (后续提出框架理论 -见Slide 2)

•关键会议/机构提及: 达特茅斯会议 (1956, AI起点), MIT AI Lab, CMU, Stanford AI Lab (早期研究中心)

符号主义的发展历程

（二）黄金时代：专家系统与知识表示 (1970s - 1980s)

- **核心进展:** 知识工程 (Knowledge Engineering) -> 专家系统 (Expert Systems)
- **代表人物及其贡献 (一一对应):**
  - **人物:** 爱德华·费根鲍姆 (Edward Feigenbaum)
    - **机构:** 斯坦福大学 (Stanford University - Heuristic Programming Project, HPP)
    - **技术/理论:** 专家系统 (倡导与领导, "知识就是力量"), DENDRAL 项目 (首批成功专家系统之一)
  - **人物:** 布鲁斯·布坎南 (Bruce Buchanan)
    - **机构:** 斯坦福大学 (Stanford University - HPP)
    - **技术/理论:** DENDRAL / MYCIN (核心开发, 知识获取与推理方法)

  人物: 马文·明斯基 (Marvin Minsky)
    - **机构:** MIT 人工智能实验室 (MIT AI Lab)
    - **技术/理论:** 框架理论 (Frame Theory - 知识表示结构化方法)
- **关键技术/公司提及:** 基于规则系统 (Rule-Based Systems), Prolog (逻辑编程), DEC (XCON应用), Symbolics, LMI (AI硬件/软件公司)
- **挑战重申:** 知识获取瓶颈, 脆弱性, 常识问题, 框架问题

符号主义的发展历程

（三）演变与融合：新方向 (1990s - 至今)

- **主要趋势:** 反思与融合 (连接主义), Web化知识表示, 追求可解释性
- **代表人物/概念及其贡献 (一一对应，部分为概念对应):**
  - **人物:** 蒂姆·伯纳斯-李 (Tim Berners-Lee)
    - **机构:** W3C (World Wide Web Consortium) / MIT
    - **技术/理论:** 语义网 (Semantic Web - 愿景提出与标准推动, RDF/OWL基础)
  - **概念: 知识图谱 (Knowledge Graphs)**
    - **机构:** Google, Microsoft, Meta (主要应用与推动者) + 学术界
    - **技术/理论:** 大规模结构化知识表示与应用 (搜索引擎优化, 问答系统等)
  - **概念: 混合智能系统 (Hybrid AI Systems)**
    - **机构:** 全球各大AI研究实验室 (e.g., MIT, Stanford, DeepMind 等)
    - **技术/理论:** 神经符号计算 (Neuro-Symbolic approaches - 结合学习与推理)
- **其他相关技术:** 约束求解 (Constraint Solving), 自动规划 (Automated Planning), 描述逻辑 (Description Logics - OWL基础)

人工智能研究方法之一：符号主义

基本特征：

（1）立足于逻辑运算和符号操作，适合于模拟　　人的逻辑思维过程，解决需要逻辑推理的复杂问题。

（2）知识可用显示的符号表示，在已知基本规则的情况下，无需输入大量的细节知识。

（3）模块化结构，当个别事实发生变化时，易于修改。

（4）能与传统的符号数据库进行连接。

（5）可对推理结论进行解释，便于对各种可能性进行选择。

主要缺陷：

可以解决逻辑思维，但对于形象思维难于模拟信息表示成符号后，并在处理或转换时，信息有丢失的情况。

人工智能的研究方法之二：联结主义

联结主义：以网络连接为主的连接机制方法。又称为自下而上和联结主义，属于非符号处理范畴。在现实中，人们并不仅仅依靠逻辑推理来求解问题，有时非逻辑推理还其着非常重要的作用。

联结主义的研究原理是神经网络及神经网络间的连接机制与学习算法，起源于仿生学，特别是人脑模型的研究，其学派代表人物是卡洛克、皮茨、Hopfield、鲁梅尔哈特等。

联结主义的发展历程

（一）萌芽与奠基 (1940s - 1960s)

1.人物: 沃伦·麦卡洛克 & 沃尔特·皮茨 (W. McCulloch & W. Pitts)

1. **机构:** 伊利诺伊大学 / MIT (U. Illinois / MIT)
2. **贡献:** MP 形式神经元模型 (提出简化的神经元数学模型)

2.人物: 弗兰克·罗森布拉特 (Frank Rosenblatt)

1. **机构:** 康奈尔航空实验室 (Cornell Aeronautical Laboratory)
2. **贡献:** 感知机 (Perceptron - 首个可学习的神经网络模型及算法)

3.人物: 马文·明斯基 & 西摩尔·派珀特 (M. Minsky & S. Papert)

1. **机构:** MIT
2. **贡献:** 《感知机》专著 (系统分析并指出单层感知机局限性，影响深远)

联结主义的发展历程

（二）蛰伏与探索 (1970s - Early 1980s)

1.人物: 保罗·韦尔博斯 (Paul Werbos)

1. **机构:** 哈佛大学 (Harvard University - PhD Thesis)
2. **贡献:** 反向传播算法 (BP) 的早期思想提出 (虽未立即普及)

2.人物: 福岛邦彦 (Kunihiko Fukushima)

1. **机构:** NHK (日本放送协会) 科学技术研究实验室
2. **贡献:** 新认知机 (Neocognitron - 具层级结构，启发CNN)

3.人物: 约翰·霍普菲尔德 (John Hopfield)

1. **机构:** 加州理工学院 / 贝尔实验室 (Caltech / Bell Labs)
2. **贡献:** 霍普菲尔德网络 (Hopfield Network - 递归网络，用于联想记忆与优化)

联结主义的发展历程

（三）第二次浪潮：反向传播与关键结构 (Mid 1980s - 1990s)

1.人物: 辛顿, 鲁梅哈特, 威廉姆斯 (G. Hinton, D. Rumelhart, R. Williams)

1. **机构:** UCSD PDP 小组 / 多伦多大学等 (UCSD PDP Group / U. Toronto, etc.)
2. **贡献:** 反向传播算法的重新发现与系统阐述 (使其广泛应用)

2.人物: 杨立昆 (Yann LeCun)

1. **机构:** AT&T 贝尔实验室 / 纽约大学 (AT&T Bell Labs / NYU)
2. **贡献:** 卷积神经网络 (CNN) 及 LeNet (成功应用于手写数字识别)

3.人物: 赛普·霍克赖特 & 于尔根·施密德胡伯 (S. Hochreiter & J. Schmidhuber)

1. **机构:** 慕尼黑工业大学 / IDSIA (TU Munich / IDSIA)
2. **贡献:** LSTM (长短期记忆网络 - 解决RNN梯度消失/爆炸问题)

联结主义的发展历程

（四）深度学习突破 (Approx. 2006 - 2012)

1.人物: 杰弗里·辛顿 (Geoffrey Hinton)

1. **机构:** 多伦多大学 / 谷歌 (U. Toronto / Google)
2. **贡献:** 深度信念网络 (DBN), ReLU 激活函数 (与Nair), Dropout (与 Srivastava等) - (关键技术突破)

2.人物: 约书亚·本吉奥 (Yoshua Bengio)

1. **机构:** 蒙特利尔大学 / MILA (U. Montreal / MILA)
2. **贡献:** 深度学习基础理论 (表征学习, 深度自编码器等)

3.人物: 吴恩达 (Andrew Ng)

1. **机构:** 斯坦福大学 / 谷歌大脑 (Stanford University / Google Brain)
2. **贡献:** 大规模无监督学习, 推动GPU在深度学习中的应用 (如 "猫" 识别实验)

联结主义的发展历程

（五）规模化、数据集与架构创新 (Approx. 2012 - 2018)

1. **人物:** 李飞飞 (Li Fei-Fei)
    1. **机构:** 斯坦福大学 (Stanford University) / (后任职 Google Cloud AI/ML 等)
    2. **贡献:** ImageNet (创建大规模标注图像数据集, 极大推动计算机视觉发展)
2. **人物:** Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton
    1. **机构:** 多伦多大学 (University of Toronto)
    2. **贡献:** AlexNet (赢得2012 ImageNet竞赛, 标志深层CNN时代来临)
3. **人物:** 何恺明 (Kaiming He) 及团队
    1. **机构:** 微软亚洲研究院 / Meta AI (MSRA / FAIR)
    2. **贡献:** ResNet (残差网络 - 解决深度网络退化问题, 实现数百甚至上千层网络训练)
    3. **(提及):** Ian Goodfellow (GANs 发明者 @ U. Montreal)

联结主义的发展历程

（六）现代纪元：Transformer、基础模型及未来 (Approx. 2017 - 至今 [2025])

1. **人物:** Ashish Vaswani 等 (Google Brain / Research 团队)
    1. **机构:** 谷歌 (Google)
    2. **贡献:** Transformer 架构 ("Attention Is All You Need" - 彻底改变NLP, 奠定大模型基础)
2. **人物/团队:** OpenAI 研究员/工程师
    1. **机构:** OpenAI
    2. **贡献:** GPT 系列 (引领大型语言模型发展), DALL-E/Sora (文生图/视频), ChatGPT (交互式AI)
3. **人物/团队:** DeepMind 研究员/工程师
    1. **机构:** DeepMind / 谷歌 (DeepMind / Google)
    2. **贡献:** AlphaGo (围棋突破), AlphaFold (蛋白质结构预测), Gemini (多模态大模型)
    3. **(提及):** Diffusion Models (扩散模型 - 高质量生成), 多模态学习, AI伦理与对齐 (Multimodality, AI Ethics & Alignment)

人工智能的研究方法之二：联结主义

基本特征：

(1) 通过神经元之间的并行协作实现信息处理, 处理过程具有并行性, 动态性, 全局性

(2) 可以实现联想的功能, 便于对有噪声的信息进行处理

(3) 可以通过对神经元之间连接强度的调整实现学习和分类等

(4) 适合模拟人类的形象思维过程

(5) 求解问题时, 可以较快的得到一个近似解。

主要缺陷：(1) 不适合于解决逻辑思维；(2) 体现结构固定和组成方案单一的系统也不适合多种知识的开发。

人工智能研究方法之三：行为主义

行为主义又称为进化主义或控制论学派，是基于控制论“动作感知”型控制系统的人工智能学派，属于非符号处理方法行为

基本观点：

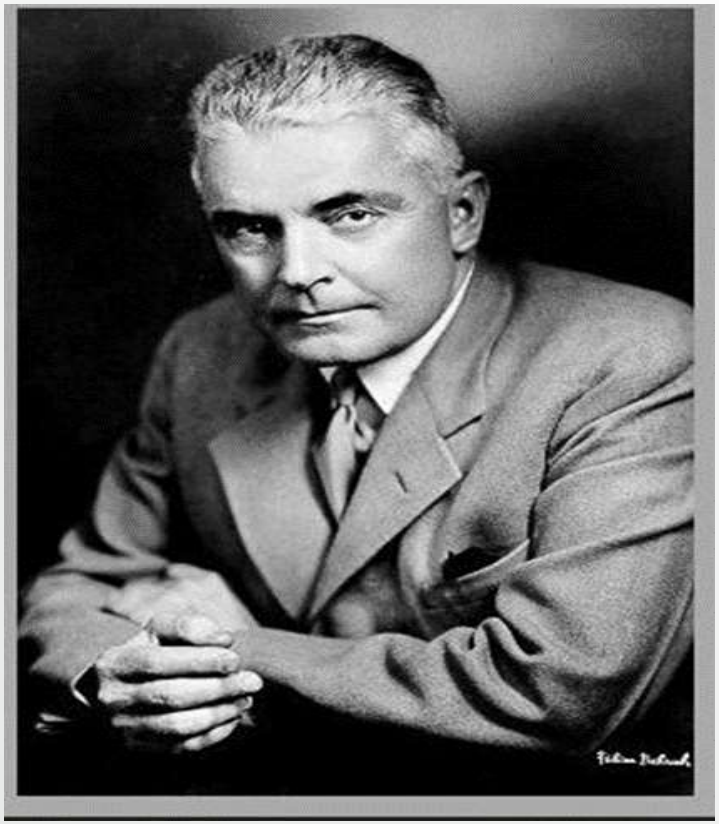
（1）知识和形式化表达和模型化方法是人工智能的重要障碍之一

（2）智能取决于感知和行动，应直接利用机器对环境作用后，环境对作用的响应为原形

（3）智能行为只能现实在世界中与周围环境交互作用而表现出来

（4）人工智能可以像人类智能一样逐步进化，分阶段发展和增强。

人工智能研究方法之三：行为主义



行为主义的AI观：认为人工智能起源于控制论，智能取决于感知和行动(所以被称为行为主义)它不需要知识、不需要表示、不需要推理。

代表人物：布鲁克(R.A. Brooks)：人的本质能力是在动态环境中的行为能力、对外界事物的感知能力、维持生命和繁衍生息的能力，正是这些能力对智能的发展提供了基础，因此智能行为只能在与现实世界的环境交互作用中表现出来，这似乎符合达尔文的进化论，即人工智能也会像人类智能一样通过逐步进化而实现(所以称为进化主义)，而不需要有知识表示和知识推理。

人工智能研究方法之三：行为主义

Brooks的代表性成果是他所研制的6足机器虫。Brooks认为要求机器人像人一样去思维太困难了，在做一个像样的机器人之前，不如先做一个像样的机器虫，由机器虫慢慢进化， 或许可以做出机器人。于是他在MIT的AI实验室研制成功了一个由150个传感器和23个执行器构成的像煌虫一样能做6足行走的机器人实验系统。这个机器虫虽然不具有像人那样的推理、规划能力，但其应付复杂环境的能力却大大超过了原有的机器人, 在自然(非结构化)环境下，具有灵活的防碰撞和漫游行为 。

目前这一观点尚未形成完整的理论体系，有待进一步研究，但由于它与人们的传统看法完全不同，因而引起了人工智能界的注意。

行为主义的发展历程

（一）行为主义启发与早期探索 (1950s - 1970s)

1. **人物:** 威廉·格雷·沃尔特 (W. Grey Walter)
    1. **机构:** Burden 神经学研究所 (英国) (Burden Neurological Institute, UK)
    2. **贡献:** "机器龟" (Machina Speculatrix) - 展示基于简单反馈回路的复杂自主行为
2. **人物:** 马文·明斯基 (Marvin Minsky)
    1. **机构:** MIT
    2. **贡献:** SNARC (1951) - 早期尝试构建通过模拟"奖惩"进行学习的机器
3. **人物:** 唐纳德·米奇 (Donald Michie)
    1. **机构:** 爱丁堡大学 (University of Edinburgh)
    2. **贡献:** MENACE (火柴盒学习机) - 使用物理奖惩机制学习井字棋的早期强化学习实例

(注：此阶段受控制论思想影响，强调系统与环境的互动与反馈)

行为主义的发展历程

（二）强化学习 (RL) 基础理论 (1980s - Early 1990s)

1.人物: 理查德·贝尔曼 (Richard Bellman)

1. **机构:** 兰德公司 (RAND Corporation)
2. **贡献:** 动态规划 (Dynamic Programming) & 贝尔曼方程 (为RL中的价值函数和最优策略提供了数学基础)

2.人物: 安德鲁·巴托 & 理查德·萨顿 (Andrew Barto & Richard Sutton)

1. **机构:** 马萨诸塞大学阿默斯特分校 (University of Massachusetts Amherst)
2. **贡献:** RL 问题的现代形式化描述, Actor-Critic 架构的早期工作

3.人物: 罗纳德·霍华德 (Ronald Howard)

1. **机构:** 斯坦福大学 (Stanford University)
2. **贡献:** 马尔可夫决策过程 (MDPs) 的系统化研究 (成为RL的标准数学框架)

行为主义的发展历程

（三）关键 RL 算法与早期成功 (1990s)

1.人物: 克里斯托弗·沃特金斯 (Christopher Watkins)
   1. **机构:** 剑桥大学 (University of Cambridge - PhD Thesis)
   2. **贡献:** Q-Learning 算法 (里程碑式的无模型 (model-free) 强化学习控制算法)

2.人物: 理查德·萨顿 (Richard Sutton)
   1. **机构:** 马萨诸塞大学 / 阿尔伯塔大学 (UMass Amherst / University of Alberta)
   2. **贡献:** 时间差分学习 (Temporal Difference, TD Learning) (核心的无模型预测和控制方法)

3.人物: 杰拉尔德·特索罗 (Gerald Tesauro)
   1. **机构:** IBM T.J. Watson 研究中心 (IBM Research)
   2. **贡献:** TD-Gammon (结合TD学习与神经网络，在双陆棋上达到世界顶级水平的程序)

(提及: 罗德尼·布鲁克斯 (Rodney Brooks) @ MIT - 包容体系结构 (Subsumption Architecture), 行为式机器人学的代表)

行为主义的发展历程

（四）深度强化学习 (Deep RL) 时代 (Approx. 2013 - 至今 [2025])

1.人物: 沃洛迪米尔·姆尼赫 (Volodymyr Mnih) 及 DeepMind 团队
   1. **机构:** DeepMind / 谷歌 (DeepMind / Google)
   2. **贡献:** DQN (深度Q网络) - 结合深度CNN与Q-Learning, 从像素输入玩转Atari游戏

2.人物: 大卫·西尔弗 (David Silver), 杰米斯·哈萨比斯 (Demis Hassabis) 及 DeepMind 团队
   1. **机构:** DeepMind / 谷歌 (DeepMind / Google)
   2. **贡献:** AlphaGo / AlphaZero - 结合蒙特卡洛树搜索、深度学习和强化学习，精通围棋等棋类游戏

3.人物/团队: OpenAI 研究员/工程师
   1. **机构:** OpenAI
   2. **贡献:** PPO (近端策略优化) 算法, 在 Dota 2 等复杂游戏中取得成功, 机器人灵巧操控