# 从AlphaGo到MuZero：技术突破与核心思想演进分析报告

## 1. 引言：AlphaGo系列——人工智能的范式革新

DeepMind开发的AlphaGo系列代表了人工智能领域的里程碑式成就，在围棋、象棋和将棋等复杂策略游戏中展现出超越人类的水平，并最终发展到能够从零开始学习游戏规则。这一系列进展深刻影响了人工智能研究，尤其是在强化学习领域 1。2016年，AlphaGo战胜世界冠军李世石的事件，成为了一个分水岭，展示了人工智能在以往被认为需要人类直觉和创造力的领域所具备的潜力 1。

围棋以其深奥的复杂性著称，其搜索空间之巨大（甚至超过宇宙中的原子数量），使得传统的“暴力”搜索方法在围棋上是不可行的 1。其复杂性主要源于难以构建有效的评估函数以及巨大的分支因子 5。此前，专家们预测人工智能至少还需要十年才能攻克围棋 6。

本报告旨在追溯从AlphaGo依赖人类数据，到MuZero能够“白板学习”（*tabula rasa*）游戏动态的演进过程，重点阐述关键的技术突破和思想转变 3。

AlphaGo系列不仅仅是解决了复杂的游戏，它更迫使我们重新评估了以往被认为是人类独有特质的“直觉”和“创造力”。AlphaGo在与李世石的对弈中下出的“第37手”，被职业棋手评价为“不像是人类的招法”且“充满美感”，正是这一点的体现 1。观察AlphaGo的棋局，可以发现它的一些招法最初令顶尖人类棋手费解，甚至被认为是失误，但事后证明这些招法具有深远的战略意义。人类在围棋中的直觉建立在长年累月的经验积累和模式识别之上。AlphaGo通过海量的（自我）对弈，发展出了自己独特的“直觉”，这种直觉基于人类尚未发现或未曾重视的模式。这表明，“直觉”可能是一种高度复杂的模式识别和预测形式，可以通过机器学习获得，而非人类专属的某种玄奥能力。这一认知挑战了以人类为中心的高级认知观点，并为人工智能在其他人类直觉可能受限或存在偏见的复杂领域中发现新颖解决方案开辟了道路。

AlphaGo的成功，如同一次“斯普特尼克时刻”，极大地推动了深度强化学习（DRL）的研究和投入，并展示了其超越学术研究的巨大潜力 1。AlphaGo的胜利吸引了公众和科学界的广泛关注 2。在此之前，DRL虽然在某些领域（如雅达利游戏）取得了一些成功，但围棋被视为难度远超以往的“重大挑战” 6。通过DRL技术攻克围棋，为这些技术的强大能力和可扩展性提供了强有力的证明，从而吸引了更多的资金投入和人才涌入，并推动了DRL在更广泛问题上的应用。这种加速发展的势头，直接促成了后续AlphaZero和MuZero等系统的快速迭代，并将DRL的原理应用于蛋白质折叠（AlphaFold）等科学领域 11。

## 2. AlphaGo：深度强化学习在围棋领域的开拓者

AlphaGo的核心思想在于通过结合深度神经网络（DNNs）进行走子预测和局面评估，并辅以先进的蒙特卡洛树搜索（MCTS）算法，从而掌握围棋 1。其关键在于将围棋巨大的搜索空间缩小到可管理的程度 7。

### 技术突破

AlphaGo的技术突破主要体现在其独特的双神经网络架构以及与MCTS的深度融合。

#### 双神经网络架构

AlphaGo采用了两个主要的神经网络：策略网络和价值网络。

* **策略网络 (pσ​ 和 pρ​):**
  + 该网络用于预测下一步棋的概率分布 1。
  + 它是一个13层的卷积神经网络（CNN）6。
  + **监督学习（SL）策略网络 (pσ​):** 通过学习约3000万个人类专家棋局（来自KGS围棋服务器数据集）进行训练，以模仿专家的下法。在预测人类专家走子上达到了约57%的准确率，显著优于当时最先进的44%的水平 3。SL策略网络通过为MCTS提供候选棋步的先验概率，引导搜索偏向人类认为有前景的区域 6。
  + **强化学习（RL）策略网络 (pρ​):** 该网络以SL策略网络的权重作为初始值，然后通过与自身先前版本进行对弈（自对弈）来不断改进。RL策略网络最终变得比SL策略网络更强大，在与SL策略网络的对弈中胜率超过80% 6。
* **价值网络 (vθ​):**
  + 该网络用于评估给定棋局的胜率，输出一个介于[-1, +1]之间的标量值 1。
  + 其架构与策略网络类似，但只有一个输出神经元 13。
  + 价值网络通过对RL策略网络自对弈产生的大量棋局（3000万个不同局面，每局只取一个以避免相关性）的结果进行回归训练 6。由于围棋局面评估被认为极其困难，价值网络的引入是一项关键创新 7。
* **快速走子（Rollout）策略 (pπ​):** 这是一个更简单、更快速的线性Softmax策略（每步耗时2µs，而SL策略网络为3ms），用于MCTS模拟（rollout）阶段快速将棋局推演至终局，尽管其准确率较低（24.2%）6。

#### 蒙特卡洛树搜索（MCTS）集成

MCTS在实战中负责协调策略网络和价值网络 7。

* **选择（Selection）:** MCTS通过选择能够最大化 Q(s,a)+u(s,a) 的动作来遍历搜索树。其中，Q(s,a) 是动作价值，u(s,a) 是一个探索奖励项，该奖励项与SL策略网络提供的先验概率 P(s,a) 成正比，与访问次数 N(s,a) 成反比 6。
* **扩展（Expansion）:** 当到达叶节点时，会为新的合法走子创建子节点，并使用SL策略网络的先验概率 P(s,a) 初始化这些新边 13。
* **评估（Evaluation）:** 叶节点的评估结合了价值网络 vθ​(sL​) 的输出和使用快速走子策略 pπ​ 模拟到游戏结束的结果 zL​。一个混合参数 λ 用于平衡这两者（实验发现 λ=0.5 效果最佳）6。这种混合评估比单独使用任一组件更为稳健。
* **反向传播（Backpropagation）:** 根据评估结果更新路径上所有边的动作价值 Q(s,a) 和访问次数 N(s,a) 13。
* 最终的落子决策基于从根节点出发被访问次数最多的分支 14。

### 学习过程

AlphaGo的训练过程分为几个阶段：

1. **阶段一：监督学习（SL）:** 训练SL策略网络 pσ​ 以模仿人类专家走子，并训练快速走子策略 pπ​ 6。
2. **阶段二：强化学习（RL）:** 通过自对弈训练RL策略网络 pρ​，使其在 pσ​ 的基础上进一步提升 6。
3. **阶段三：价值网络训练:** 使用RL策略网络自对弈产生的数据集训练价值网络 vθ​ 6。

AlphaGo的核心思想在于利用深度神经网络捕捉人类专家的下棋方式（策略）和评估局面优劣（价值）的能力，然后将这些能力与MCTS相结合以进行高效搜索 7。

AlphaGo最初对人类专家数据的依赖（SL策略网络）扮演了关键的“脚手架”角色，引导强化学习过程走向广阔搜索空间中更有希望的区域。若没有这种初步引导，对于围棋这样复杂的游戏，在当时的技术条件下，从零开始的强化学习可能会效率低下或难以稳定。AlphaGo的训练流程明确地始于对3000万个人类棋步的监督学习 7。SL策略随后被用于初始化RL策略并指导MCTS的探索 13。围棋的搜索空间极其庞大，随机探索对于强化学习而言效率极低。人类棋局代表了一个经过筛选的“好”（至少是合理）策略的数据集。SL阶段提供了一个强大的先验知识，有效地修剪了搜索空间，并为RL提供了一个良好的起点。这使得后续的RL阶段更易于处理且更有效率，使AlphaGo能够发现超越直接模仿但仍植根于合理棋局的策略。这凸显了一种解决复杂AI问题的实用方法：利用现有（人类）知识来引导学习，然后使用自我改进机制来超越初始知识。这是在AlphaGo Zero的“纯强化学习”方法变得可行之前的一个关键步骤。

值得注意的是，无论是策略网络、价值网络还是MCTS中的快速走子，单独来看都不是完美的。SL策略网络的准确率仅为57% 6，价值网络提供的是近似评估 13，而快速走子的准确率更低，仅约24% 13。AlphaGo的强大之处在于MCTS框架能够稳健地结合这些不完美但互补的组件。MCTS以其通过多次模拟聚合噪声评估以找到优质走子的能力而闻名。策略网络引导搜索的宽度，价值网络减少搜索的深度，而快速走子则提供了另一种评估信号。AlphaGo的架构证明，一个精心设计的搜索算法（MCTS）可以有效地整合多个各自不完美的启发式评估函数（策略网络、价值网络、快速走子），从而产生远超任何单个组件的集体智能。混合评估（λ=0.5）表现最佳的事实也支持了这一点 14。这为复杂领域的人工智能设计提供了一个原则：专注于构建能够利用多样化、可能带有噪声的信息源的稳健集成机制（如MCTS），而不是努力追求单一完美的整体模型。这也预示了后来关于集成方法和多组件AI系统的工作。

此外，AlphaGo在与李世石的比赛中据报道使用了1920个CPU和280个GPU 5，这突显了其所需的巨大计算资源。这种高昂的成本可能也成为了后续AlphaGo Zero追求算法效率的驱动力。AlphaGo（李世石版本）使用了大量的分布式计算资源 5，其训练也耗费了大量时间和资源 13。高计算成本限制了其可及性、可扩展性以及研发中的快速迭代。AlphaGo的成功虽然具有开创性，但也强调了对更高效算法的需求，这些算法可以用更少的计算资源达到相似或更好的性能。这种实际限制可能成为AlphaGo Zero在架构和算法上进行简化的一个关键驱动因素（例如，统一网络、取消快速走子）。在AI领域，追求算法效率与性能并重是一个反复出现的主题。突破往往不仅在于实现某种能力，还在于以更节省资源的方式实现，这扩大了适用性并加速了进一步的发展。

## 3. AlphaGo Zero：从第一性原理掌握围棋

AlphaGo Zero的核心思想是实现“白板学习”（*tabula rasa*），即完全通过自对弈强化学习，不依赖任何人类棋谱数据、监督信息或除基本规则之外的领域知识，来达到超越人类的围棋水平 3。这种方法旨在摆脱人类知识的局限，发现全新的策略 10。

### 技术突破

AlphaGo Zero的技术突破体现在其纯粹的自学习能力、统一的网络架构以及简化的学习算法。

#### *Tabula Rasa* 学习

* 系统从一个除了围棋规则外一无所知的神经网络开始 17。
* 训练过程中不使用任何人类棋谱数据 3。
* 仅使用棋盘上的黑白子作为输入特征（无人工设计的特征）17。

#### 统一的策略与价值网络架构

* 将策略网络和价值网络整合进一个单一的深度神经网络（采用ResNet残差网络架构，包含两个输出头）3。
* 这个单一网络 fθ​(s) 以棋盘状态 s 作为输入，同时输出走子概率 **p**（策略头）和一个标量值 *v*（价值头，预测胜率）17。
* 这种统一架构使得训练和评估更为高效 3。

#### 简化的MCTS与学习算法

* MCTS搜索完全依赖这个单一神经网络进行局面评估和走子采样 17。
* **取消快速走子（Rollouts）:** 与AlphaGo不同，AlphaGo Zero在局面评估时不使用快速、随机的棋局推演至终局。它完全依赖其高质量神经网络的价值输出 17。
* **强化学习循环:**
  1. 当前的神经网络指导MCTS进行自对弈。
  2. MCTS搜索输出改进后的走子概率 **π**。
  3. 对弈的最终结果 *z*（赢/输）作为价值网络的目标。
  4. 更新神经网络参数 θ，以最小化其预测价值 *v* 与对弈结果 *z* 之间的误差，并最大化其走子概率 **p** 与MCTS搜索概率 **π** 的相似性 21。
  5. 此过程不断迭代，网络能力越来越强，从而产生更高质量的自对弈棋局和更准确的网络更新 17。

#### 学习效率与性能的显著提升

* 学习速度极快：仅用3天自对弈训练就超越了AlphaGo Lee（击败李世石的版本），并以100:0的比分获胜 3。
* 经过40天训练后，其表现优于AlphaGo Master（曾击败顶尖职业棋手）17。
* 以远少于初代AlphaGo的计算资源和更简洁的架构，实现了更优越的性能 3。
* 发现了全新的围棋策略和知识，在几天内超越了人类数千年积累的知识 10。

AlphaGo Zero通过“白板学习”超越了依赖人类数据的AlphaGo，这验证了一个重要假设：人类知识在引导AI的同时，也可能成为其发展的上限，因为它可能将人类的偏见或次优策略传递给AI 10。AlphaGo Zero在不使用人类数据的情况下以100:0击败AlphaGo Lee 17，DeepMind也指出它“不再受人类知识极限的束缚” 10。人类专家的棋局虽然水平很高，但并非绝对最优，并且包含了固有的偏见或历史上偏好的下法，这些下法可能并非全局最佳。通过从零开始学习，AlphaGo Zero得以探索围棋的整个策略空间，不受人类先入为主观念的阻碍，从而发现了全新的、更强大的下棋方式。一个使用人类棋谱训练的版本虽然初期学习更快，但长期表现反而更差 18，这一点有力地支持了上述观点。这对AI发展具有深远影响，表明对于那些最优解未知的足够复杂的问题，纯粹的自对弈强化学习可能比基于（即便是专家）人类数据的监督学习更为强大，有潜力解锁真正新颖的解决方案。

AlphaGo Zero架构的简化（统一网络、无快速走子、无人工特征）不仅仅是为了降低复杂性，更是其卓越学习效率和性能的关键促成因素。AlphaGo Zero使用一个神经网络代替了两个，取消了快速走子，并且仅使用原始棋盘状态作为输入 17。同时，它的学习速度更快，计算量也更少 3。分离的网络可能产生冲突的梯度或学习冗余的特征。快速走子增加了计算开销和噪声。人工特征需要领域专业知识，并且可能遗漏关键信息或引入偏见。统一的网络使得学习信号更加直接和一致。取消快速走子迫使网络发展出对局面评估更强的内在理解。仅使用原始特征则迫使网络从第一性原理学习所有相关表示。这些简化创造了一个更专注、更强大的学习目标。这倡导了一种理念，即算法的优雅和找到正确的表示抽象层次，可能比简单地增加更多组件或特征更具影响力。当核心学习算法足够强大时，这正是“少即是多”的体现。

DeepMind的David Silver提出的“自对弈能够提供一个‘恰到好处水平’的对手” 20 这一观点至关重要。系统总是面临挑战但又不会被完全压制，从而形成了一个最优的学习曲线。AlphaGo Zero通过与自身对弈进行学习，从完全随机的下法开始 17，其性能逐步迭代提升 17。在学习过程中，如果对手太弱，则学不到太多东西；如果对手太强，智能体很少成功，获得的学习信号也很差。自对弈内在地提供了一种课程，其中对手的强度与智能体当前的能力相匹配。随着智能体的进步，其对手（自身）也在进步，确保了持续有效的学习梯度。这种动态难度调整是自对弈强化学习一个强大且固有的特性。这突显了机器学习中自生成课程的力量。对于复杂任务，设计一个最优的外部课程可能非常困难。自对弈为智能体提供了一种创建其自身完美定制学习环境的方式，这一概念在人工智能教育和技能获取方面具有广泛的适用性。

## 4. AlphaZero：跨越多领域的通用泛化

AlphaZero的核心思想是创建一个单一的、通用的算法，该算法能够沿用AlphaGo Zero的架构和学习范式，从零开始在多种具有挑战性的游戏中（围棋、象棋、将棋）达到超越人类的水平 1。这表明“白板学习”的强化学习方法可以推广到单一游戏之外 30。

### 技术突破

AlphaZero的突破在于其通用的强化学习框架以及针对多游戏泛化的适应性调整。

#### 通用强化学习框架

* 对围棋、象棋和将棋均采用了相同的基本算法、网络架构（带有策略头和价值头的深度CNN）以及MCTS方法 25。
* 完全通过自对弈学习，仅给定各游戏的基本规则。除了规则之外，不使用任何特定于游戏的人类数据或人工设计的特征 25。

#### 相对于AlphaGo Zero的架构与算法调整/泛化

* **处理平局:** AlphaZero的价值头被调整为预测期望结果（赢/输/平局），而不仅仅是赢/输概率，以适应象棋和将棋这类可能出现平局的游戏 18。
* **无特定游戏对称性利用:** 与AlphaGo Zero（利用围棋的旋转/反射对称性进行数据增强）不同，AlphaZero不使用特定于游戏的对称性，这使得算法更具通用性 25。
* **持续学习:** AlphaZero维护一个单一的神经网络，该网络通过最新参数生成的自对弈游戏进行持续更新，而不是像AlphaGo Zero那样采用迭代评估新玩家与当前最佳玩家的模式 18。
* **共享超参数:** MCTS搜索在所有游戏中使用了相同的超参数，未进行特定于游戏的调整（除了探索噪声等少数例外）25。
* **输入表示:** 神经网络的输入根据不同游戏进行了调整（例如，象棋为8x8棋盘、棋子类型、重复局面计数器）26。

#### 性能表现

* 在所有三种游戏中均迅速达到超人水平：
  + 象棋：经过9小时训练后超越Stockfish（顶级象棋引擎）27。
  + 将棋：经过12小时训练后超越Elmo（顶级将棋引擎）27。
  + 围棋：经过34小时训练后超越AlphaGo Zero（击败AlphaGo Lee的3天版本），在与某个版本的AlphaGo Zero的对弈中以60:40获胜 25。 38 Elo为4987，而AGZ 40层为5185)。
* 尽管每秒评估的局面数量远少于传统引擎（例如，AlphaZero在象棋中为8万，而Stockfish为7000万），但AlphaZero通过其深度神经网络更具选择性地聚焦于有希望的变着，从而弥补了这一不足 25。
* 在象棋等游戏中发展出非传统且富有创造性的棋风 27。

AlphaZero在三种不同且高度复杂的游戏中，仅对核心算法进行微小改动便取得成功，这标志着向通用人工智能问题解决能力迈出了重要一步。它表明，通过深度神经网络和MCTS从自对弈中学习的基本原理并不仅限于围棋。AlphaZero对所有三种游戏使用了“相同的算法和网络架构” 30。这些游戏具有不同的规则、棋盘大小、棋子类型和战略细微差别。传统的游戏AI需要为每个游戏构建高度专业化的人工智能引擎。AlphaZero能够用一种通用方法掌握所有这些游戏，这意味着它学习了这些游戏共有的战略博弈的基本原理，而不仅仅是特定于游戏的模式。网络学会了表示和推理不同的状态空间和动作类型。这是验证通用人工智能追求的关键一步。它表明，一个单一的学习系统，只要有足够的经验（即使是自我生成的），就能够适应各种复杂的基于规则的环境，超越了狭义人工智能的范畴。

AlphaZero尽管搜索的局面数量比Stockfish等专业引擎少几个数量级，却能击败它们，这表明其神经网络捕捉到的博弈知识（策略和价值）比这些引擎中数十年来人类工程师积累的启发式规则更为丰富和有效。AlphaZero在象棋中每秒搜索的局面远少于Stockfish（例如8万对7000万），但仍能决定性地获胜 25。传统引擎依赖高度优化的搜索（alpha-beta剪枝）和精心制作的评估函数。AlphaZero的搜索（MCTS）则由其神经网络的策略和价值输出引导。神经网络必须提供远为优越的指导——无论是在识别有希望的走子（策略）方面，还是在评估局面（价值）方面——才能弥补搜索深度的不足。这意味着网络学习到了对游戏战略格局更为细致和准确的“理解”。这挑战了长期以来依赖人类专家将知识提炼为AI启发式规则的范式。它表明，在某些领域，深度学习可以比人类编码更有效地自动发现和表示复杂知识，特别是当与自对弈结合以产生海量经验时。

与AlphaGo Zero的迭代“竞赛”模型相比，AlphaZero转而采用单一、持续更新的网络 18，这代表了一种更敏捷、可能也更快速的AI改进方式。AlphaZero持续更新其单一网络，自对弈游戏始终使用最新的参数 30。而AlphaGo Zero则有明确的训练迭代，之后通过比赛来决定新网络是否成为“最佳玩家” 31。具有离散评估点的迭代模型可能会在整合改进方面引入延迟。如果一个新网络只是略有改进，或者在某些方面有所改进但在其他方面没有，它可能无法通过成为新“最佳”的门槛，从而减缓整体进展。持续更新则允许更快地整合学习到的改进。每一次自对弈都直接有助于完善单一网络，从而实现更平稳、可能也更快速地收敛到更高性能。这反映了机器学习向在线或持续学习系统发展的更广泛趋势，这些系统能够更动态地适应，这对于数据分布可能频繁变化或新信息不断出现的实际应用至关重要。

## 5. MuZero：学习游戏规则的智能体

MuZero的核心思想是，通过从观察中学习环境动态模型，而非被明确告知规则，来掌握各种游戏（围棋、象棋、将棋以及雅达利游戏），然后利用这个学习到的模型进行规划 3。这与AlphaZero需要完美模拟器（即了解规则）相比，是一个显著的进步 3。

### 技术突破

MuZero的技术突破在于其创新的学习环境模型的方法以及基于该模型的规划能力。

#### 学习环境动态模型

MuZero学习一个模型，该模型仅预测与规划相关的环境方面 11。该模型主要由三个神经网络组件构成：

* **表示函数 (h):** 将观察（例如，棋盘状态、雅达利游戏画面像素）映射到一个隐藏状态 s0​。这个隐藏状态是一种内部学习到的表示 3。
* **动态函数 (g):** 给定前一个隐藏状态 sk−1​ 和一个假设的动作 ak​，它预测下一个隐藏状态 sk​ 和即时奖励 rk​。这使得MuZero能够“前推”其内部模型 3。
* **预测函数 (f):** 接收一个隐藏状态 sk​ 并预测策略 pk​（动作的概率分布）和价值 vk​（从此状态预期的未来回报）3。

#### 基于学习模型的规划（MCTS内部）

* MCTS在这些学习到的隐藏状态空间中进行 3。
* 在搜索过程中，当遍历一条边（模拟一个动作）时，使用动态函数 *g* 转换到下一个隐藏状态并预测奖励。在新的隐藏状态下，使用预测函数 *f* 获取策略和价值估计 36。
* 模型通过端到端的方式进行训练，以准确预测策略、价值和观察到的奖励，使其与搜索产生的改进估计相匹配 3。

#### 无需明确规则知识

MuZero不需要了解游戏规则或环境如何运作。它通过其学习到的模型隐式地学习这些知识 3。

#### 性能与通用性

* 在围棋、象棋和将棋中达到或超过了AlphaZero的性能，尽管在某些情况下（例如，围棋中每个节点的搜索计算量更少，MuZero使用16个残差块，而AlphaZero使用20个 38）3。
* 在视觉复杂的雅达利学习环境（57款游戏）中取得了最先进的成果，而基于模型的规划方法在该领域历来举步维艰 3。
* 这证明了该算法适用于更广泛的环境，包括那些具有复杂视觉输入和未知动态的环境 8。

MuZero的成功表明，为了进行有效规划，人工智能系统无需学习对环境进行完美、像素级的重构。相反，学习一个仅捕捉与价值和策略预测相关的动态模型，不仅是充分的，而且效率更高。MuZero学习到的模型预测策略、价值和即时奖励，而不必是完整的观察空间 11。隐藏状态不需要重构原始观察 41。传统的基于模型的强化学习通常难以学习高保真度的世界模型，特别是对于像雅达利这样的视觉丰富环境。MuZero的突破在于，它通过让模型只学习决策所必需的内容，从而回避了这个问题。内部隐藏状态“可以自由地以任何与预测当前和未来价值及策略相关的方式来表示状态” 41。这使得模型学习任务更易于处理。这对于在复杂的现实世界场景中构建智能体是一个强有力的思想。智能体无需试图理解关于世界的一切，而是可以专注于学习一个紧凑的、与任务相关的模型，从而使学习更具可扩展性，并且对无关细节更具鲁棒性。这与人类构建抽象心智模型的方式有相似之处。

通过使用表示函数 (h) 将观察映射到抽象的隐藏状态，然后让动态函数 (g) 和预测函数 (f) 在此隐藏状态上操作，MuZero有效地将感知挑战与学习环境动态和规划的挑战分离开来。MuZero的架构明确包含一个表示函数 h(o)→s0​，随后是动态函数 g(s,a)→s′,r 和预测函数 f(s)→p,v 3。在复杂环境（例如，从像素开始的雅达利游戏）中，原始观察是高维且充满噪声的。直接从像素学习动态非常困难。表示函数学会将原始观察提炼成一个更抽象、更稳定的潜在空间。然后，动态和预测函数可以在这种更清晰、维度更低的表示上操作，从而使其学习任务更容易。这种模块化允许网络的不同部分专门化。这种架构模式在现代人工智能中具有很高的影响力，以各种形式出现，如自动编码器或学习嵌入。它允许端到端学习，同时仍受益于一定程度的功能分解，这对于解决像同时感知、建模和规划这样的多方面问题至关重要。

即使学习到的动态模型不完美，MCTS规划过程仍然可以通过迭代优化预测和探索替代方案来获得强大的性能。规划过程可以在一定程度上弥补学习模型的不足 34。研究表明，MuZero的动态网络在较长时间的模拟中可能会变得不那么准确，但MuZero仍然通过利用规划来纠正错误而有效执行 34。MCTS本身就涉及探索多个未来轨迹并对结果进行平均。搜索过程中每一步的价值和策略预测都提供持续的指导。如果学习模型在预测未来状态或奖励时出现轻微错误，搜索过程通过探索多个分支并在每个新的想象状态下使用预测函数 *f* 重新评估，通常可以恢复或找到仍然良好的替代路径。搜索并不会盲目相信学习模型的单次长期推演。这种对不完美学习模型的鲁棒性对于现实世界的应用至关重要，因为在这些应用中，完美的环境模型几乎不可能获得。它表明，将学习模型与稳健的搜索/规划算法相结合，即使学习组件并非完美无瑕，也可以产生实用的解决方案。这是构建弹性AI系统的一个关键方面。

## 6. 演进轨迹与核心思想转变

从AlphaGo到MuZero的演进，不仅是技术能力的飞跃，更体现了人工智能研究核心思想的深刻转变。

### 从依赖人类数据到“白板”自监督学习

* **AlphaGo:** 严重依赖监督学习，使用人类专家棋局进行初始策略网络训练 3。这是一个使问题易于处理的实用选择。
* **AlphaGo Zero:** 完全摆脱了对人类数据的依赖，纯粹通过自对弈和基本规则进行学习 3。这证明了AI可以在没有直接模仿的情况下超越人类知识，是一次重大的思想转变 8。
* **AlphaZero:** 将“白板学习”方法推广到多种游戏 25。
* **MuZero:** 同样采用“白板学习”，但更进一步，甚至不需要明确提供规则 27。
* **影响:** 这一演进过程展示了通过强化学习进行自监督学习的巨大威力，减少了获取大规模、有标签人类数据集的瓶颈，并使AI能够探索超越人类偏见的领域 1。

### 从特定游戏到通用游戏智能的演进

* **AlphaGo:** 专门为围棋设计和优化 5。
* **AlphaGo Zero:** 仍然专注于围棋，但其“白板学习”方式是向通用性迈出的概念性一步 8。
* **AlphaZero:** 第一个证明了单一算法能够以相同的架构和超参数掌握多种复杂棋盘游戏（围棋、象棋、将棋）的系统 8。
* **MuZero:** 进一步推广到包括雅达利游戏（视觉复杂、规则多样）以及棋盘游戏，所有这些都无需预先了解规则 3。
* **影响:** 这一轨迹标志着在通用人工智能（AGI）方面取得的重大进展，展示了更不易出错、更具跨领域适应性的算法 8。

### 神经网络架构与MCTS集成的演变

* **AlphaGo:** 分离的策略网络和价值网络（CNN）；MCTS使用快速走子并结合评估结果 3。
* **AlphaGo Zero:** 统一的策略和价值网络（ResNet架构）；MCTS完全依赖该网络，取消了快速走子 3。
* **AlphaZero:** 保持了统一的ResNet架构，并针对不同游戏的输入进行了调整；MCTS适应了平局结果 25。
* **MuZero:** 引入了更复杂的网络结构，包含表示、动态和预测函数；MCTS在学习到的隐藏状态上进行规划 3。
* **影响:** 趋势首先是向架构简化和效率提升发展（AlphaGo到AlphaGo Zero），然后是学习更复杂的内部表示（MuZero的学习模型）以处理更大的不确定性和通用性。MCTS始终是一个强大的搜索机制，但其搜索对象和引导方式在不断演变 3。

### 算法优雅性、效率以及自对弈能力的不断增强

* 每一次迭代通常在算法上变得更加优雅，并且在所达到的性能方面，其计算效率在许多方面也得到了提升 3。
* AlphaGo Zero比AlphaGo学习更快，所需数据/计算量更少 3。
* MuZero在学习模型的同时，在棋盘游戏中仍能达到与AlphaZero相当或更好的性能，并在雅达利游戏中表现出色 3。
* 通过自对弈进行强化学习的核心思想变得越来越重要和强大，证明了其产生海量“经验”并驱动学习达到超人水平的能力 3。

### 向基于（学习）模型的强化学习转变

* AlphaGo、AlphaGo Zero和AlphaZero主要属于无模型（model-free）方法，因为它们不显式学习游戏规则的转移模型来进行规划（AlphaZero被给予规则/模拟器）。它们直接学习策略和/或价值函数。
* MuZero标志着一个重要的转变，它学习环境模型（动态、奖励），然后使用这个学习到的模型进行MCTS规划。这是一种模型本身也是学习得来的基于模型的强化学习（model-based RL）3。
* **影响:** 这对于那些无法获得完美模拟器或规则未知/过于复杂难以定义的领域至关重要，为许多现实世界应用打开了大门 8。

整个AlphaGo系列的发展可以看作是对一个核心成功理念——将基于深度学习的函数逼近（用于策略/价值/模型）与强大的搜索算法（MCTS）相结合——进行迭代优化和泛化的有力证明。所有这四个系统都使用了神经网络和MCTS 3。其间的变化在于这些组件的结构、训练方式以及相互作用的方式。AlphaGo证明了强化学习+搜索的可行性。AlphaGo Zero消除了对人类数据的依赖并简化了神经网络。AlphaZero将游戏领域进行了泛化。MuZero通过学习用于搜索的模型，消除了对规则的依赖。这显示了一种科学的演进过程：建立一个强大的基线，然后系统地解决其弱点或扩展其能力。在强大的搜索框架内利用学习到的评估器的核心思想得以保留，但其实现变得越来越复杂和通用。这种迭代的、问题驱动的研究方法，建立在坚实的基础概念之上，是成功科学探索的标志，并为应对其他重大AI挑战提供了典范。

实现超人智能的主要瓶颈，从最初的“如何编码人类知识”（AlphaGo面临的挑战），转变为“如何实现从零开始高效发现知识”（AlphaGo Zero/AlphaZero），最终演变为“如何学习环境本身的构造以实现知识发现”（MuZero）。AlphaGo需要人类棋谱数据 13。AlphaGo Zero只需要规则 17。MuZero则两者都不需要 36。获取高质量的人类专家数据是有限且可能成本高昂的。对于某些问题，游戏规则是明确定义的，但对于许多现实世界场景则不然。DeepMind系统地解决了日益根本的“知识瓶颈”。首先是摆脱对明确人类策略的依赖。其次是摆脱对明确环境规则的依赖。每一步都使AI更加自主，并适用于更广泛的问题。这一轨迹指向了那些在根本上更具自我学习能力的AI系统，它们能够在真正未知或知之甚少的环境中运行，这是稳健的现实世界AI的关键要求。

虽然最初的突破集中在达到峰值性能（例如，击败冠军），但后来的发展（特别是围绕MuZero及后续研究，如Gumbel MuZero 42）显示出越来越关注样本效率（用更少的模拟/游戏进行学习）和在更多样化任务上的泛化能力。AlphaGo/AlphaGo Zero需要数百万/数十亿次的自对弈 17。MuZero扩展到雅达利，一个多样化的游戏套件 36。Gumbel MuZero明确旨在通过少量模拟提高性能 42。海量数据/计算需求限制了实际应用。真正的智能意味着能够高效学习并适应新的、未见过的情况。研究前沿正在从仅仅“AI能做X吗？”转向“AI能高效、自适应地做X吗？”这涉及到开发能够学习更好表示、更智能地探索以及更有效地迁移知识的算法。这种转变为在现实世界应用中部署AI至关重要，在这些应用中，数据可能稀缺、昂贵或非平稳，并且系统需要处理各种任务而不是单一狭隘的目标。这推动了对元学习、迁移学习和更复杂的基于模型的强化学习等领域的研究。

下表总结了AlphaGo、AlphaGo Zero、AlphaZero和MuZero之间的主要区别：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **特征** | **AlphaGo** | **AlphaGo Zero** | **AlphaZero** | **MuZero** |
| **主要目标** | 在围棋中击败人类顶尖选手 | 完全通过自对弈，从零开始掌握围棋，超越AlphaGo | 创建一个能从零开始掌握多种棋盘游戏的通用算法 | 创建一个能在不知道规则的情况下，通过学习环境模型来掌握多种游戏（包括棋盘和雅达利游戏）的通用算法 |
| **核心思想** | 结合深度神经网络（策略与价值）和MCTS，利用人类专家数据进行引导 | 完全通过强化学习自对弈，使用统一的策略和价值网络，无需人类数据 | 将AlphaGo Zero的方法泛化，用单一算法和架构掌握多种游戏 | 通过学习环境的动态模型（表示、动态、预测函数）来进行规划，无需预知规则 |
| **学习人类数据?** | 是 (用于初始策略网络训练) 7 | 否 17 | 否 25 | 否 27 |
| **学习游戏规则?** | 是 (规则硬编码) | 是 (规则硬编码) | 是 (规则硬编码) | 否 (通过与环境交互学习动态) 36 |
| **神经网络架构** | 两个独立的CNN：策略网络和价值网络 13 | 单一的ResNet，包含策略头和价值头 17 | 单一的ResNet，包含策略头和价值头，针对不同游戏调整输入 26 | 包含表示函数、动态函数和预测函数的网络架构 3 |
| **策略网络** | SL策略网络 (基于人类数据) 和 RL策略网络 (自对弈改进) 13 | 统一网络中的策略头 17 | 统一网络中的策略头 26 | 预测函数输出策略 3 |
| **价值网络** | 独立的价值网络，评估胜率 13 | 统一网络中的价值头 17 | 统一网络中的价值头，预测期望结果 (包括平局) 25 | 预测函数输出价值 3 |
| **MCTS中使用Rollouts?** | 是 (使用快速走子策略) 13 | 否 (完全依赖价值网络评估) 17 | 否 26 | 否 (在学习到的隐藏状态空间中进行规划) 41 |
| **关键创新** | 深度学习与MCTS结合；策略网络与价值网络；人类数据引导RL | 从零开始的自对弈学习；统一的策略和价值网络；无Rollouts；更高的效率和性能 3 | 单一算法掌握多种复杂棋盘游戏；处理平局；无特定游戏对称性；持续学习 25 | 学习环境模型进行规划；无需预知游戏规则；在棋盘游戏和雅达利游戏中均表现出色 27 |
| **掌握的游戏** | 围棋 5 | 围棋 17 | 围棋、象棋、将棋 25 | 围棋、象棋、将棋、雅达利游戏 27 |
| **输入特征** | 棋盘状态 + 人工特征 (早期版本) / 仅棋盘状态 (后期版本) 13 | 仅棋盘上的黑白子 17 | 针对不同游戏的棋盘表示 (例如，棋子类型、重复计数) 26 | 原始观察 (例如，棋盘状态、雅达利游戏画面) 41 |
| **泛化能力** | 针对围棋特化 | 概念上向通用性迈进，但仍专注于围棋 | 在多种棋盘游戏间展现了算法的通用性 | 在棋盘游戏和规则迥异的雅达利游戏间展现了更强的通用性，尤其是不依赖规则的能力 |

## 7. 结论：历史遗产与未来展望

AlphaGo系列的演进清晰地勾勒出人工智能，特别是深度强化学习领域，在过去十年间取得的惊人进展。

### 主要技术突破与核心思想回顾

* **AlphaGo:** 证明了深度强化学习在复杂策略游戏中取得成功的可行性。其核心在于双神经网络（策略网络和价值网络）与蒙特卡洛树搜索（MCTS）的结合，并巧妙地利用人类专家数据作为学习的起点和引导。
* **AlphaGo Zero:** 实现了“白板学习”的突破，完全摆脱了对人类数据的依赖。通过统一的策略和价值网络架构以及纯粹的自对弈强化学习，它不仅在围棋上超越了AlphaGo，更重要的是证明了AI有能力在没有人类先验知识的情况下发现并超越人类积累的智慧，同时在效率和性能上也取得了显著提升。
* **AlphaZero:** 将AlphaGo Zero的“白板学习”范式成功推广到多种不同的复杂棋盘游戏（围棋、象棋、将棋），展示了单一算法在不同规则体系下的学习能力和通用性。
* **MuZero:** 进一步拓展了AI的边界，通过学习环境的动态模型（而非被动接受规则），实现了在规则未知情况下的规划和决策。这使其能够掌握包括雅达利视频游戏在内的更广泛类型的任务，代表了向更通用、更自主的人工智能迈出的重要一步。

### 对强化学习及AI研究的深远影响

AlphaGo系列的成功对人工智能研究产生了深远的影响 1：

* **验证了深度学习与强化学习大规模结合的威力:** 它展示了当这两者结合并应用于大规模计算资源时，能够解决此前被认为难以逾越的复杂问题 1。
* **推广了自对弈作为核心学习机制:** 自对弈被证明是一种极其强大的机制，能够产生海量高质量的训练数据，并驱动AI系统达到甚至超越人类的顶尖水平 3。
* **推动了从特征工程到端到端学习的转变:** 尤其是在AlphaGo Zero之后，系统更侧重于从原始输入直接学习有效的表示、策略和价值函数，减少了对人工设计特征的依赖。
* **激发了对更通用、高效、鲁棒的强化学习算法的研究:** 该系列工作为后续研究设定了很高的标杆，并指明了提升算法通用性、样本效率和鲁棒性的重要方向。
* **甚至影响了人类玩家的策略探索:** AlphaGo系列的新颖下法启发了人类职业棋手，促使他们探索新的策略和思路 44。

### 对未来AI发展及现实应用的启示

AlphaGo系列的技术和思想正被应用于更广泛的领域：

* **科学发现:** 例如，AlphaFold利用类似原理在蛋白质折叠预测方面取得了革命性突破，极大地推动了生命科学研究 11。其他在材料科学、能源优化等领域的应用也在探索中 3。
* **未知环境下的决策:** MuZero学习环境模型的能力，对于那些缺乏完美模拟器或规则未知的现实世界问题（如机器人控制、资源管理、自主系统）尤其具有应用前景 3。
* **持续的挑战:** 尽管取得了巨大成功，但在样本效率（尤其对于极其复杂的现实任务）、安全性、可解释性、以及处理多智能体协作或部分可观察环境等方面，仍面临诸多挑战 4。

从AlphaGo到MuZero的演进，清晰地描绘了一条沿着“泛化梯度”前进的道路——从依赖人类数据的单一游戏，到不依赖人类数据的单一游戏，再到不依赖人类数据的多种游戏，最终到甚至无需规则即可适应多种环境。这个梯度本身为更广泛的AI社区提供了一个强有力的研究指南，表明追求更强的泛化能力是一个富有成效的方向。每一次的进步都证明了追求更通用AI系统的价值，并突显了需要克服的关键挑战（例如，数据依赖、规则依赖）。这种轨迹鼓励AI领域不断追问：“我们如何能让这个系统在更广泛的问题上，以更少的假设来工作？”这驱动着向更基本、更鲁棒的AI原理的创新。

此外，虽然叙事往往集中在算法突破上，但大规模计算能力的可用性（例如TPUs 18）是这些系统不可或缺的推动因素，特别是对于广泛的自对弈和神经网络训练。这表明了一种共生关系：算法思想使得能够更好地利用规模，而规模则允许探索更雄心勃勃的算法思想。所有AlphaGo系列系统都利用了大量的计算资源，包括数千个TPU用于训练和自对弈 5。训练深度神经网络和运行数百万/数十亿次MCTS模拟是计算密集型的。算法创新（例如，高效的网络架构、自对弈强化学习）被设计为可扩展的，但它们的实际可行性以及达到超人性能的速度，都取决于对大规模并行计算的访问。如果没有这些资源，这些突破不太可能以如此快的速度发生。未来AI的进步很可能继续依赖于这种相互作用。算法进展需要考虑到可扩展性，而对计算基础设施的持续投资对于推动AI所能达到的极限将是必要的，尤其是在像DRL这样复杂、数据饥渴的范式中。这也引发了关于公平获取前沿AI研究所需资源的疑问。

总而言之，AlphaGo、AlphaGo Zero、AlphaZero和MuZero的演进不仅是人工智能技术的一次次辉煌展示，更是对学习、智能和创造力本质的深刻探索。它们为构建更强大、更通用的AI系统奠定了坚实的基础，并持续激励着研究者们向着解决更复杂现实问题的目标迈进。

#### Works cited

1. AlphaGo: AI Conquers Go, Redefining Human Intelligence - Open MedScience, accessed on May 18, 2025, <https://openmedscience.com/alphago-ai-conquers-go-redefining-human-intelligence/>
2. Alphago - Lark, accessed on May 18, 2025, <https://www.larksuite.com/en_us/topics/ai-glossary/alphago>
3. Reinforcement Learning in Strategy-Based and Atari Games: A ..., accessed on May 18, 2025, <https://www.alphaxiv.org/overview/2502.10303>
4. Reinforcement Learning in Strategy-Based and Atari Games: A Review of Google DeepMind's Innovations - arXiv, accessed on May 18, 2025, <https://arxiv.org/html/2502.10303>
5. AlphaGo - Wikipedia, accessed on May 18, 2025, <https://en.wikipedia.org/wiki/AlphaGo>
6. hci.iwr.uni-heidelberg.de, accessed on May 18, 2025, <https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/36349047/report_florian_brunner.pdf>
7. AlphaGo: Mastering the ancient game of Go with Machine Learning - Google Research, accessed on May 18, 2025, <https://research.google/blog/alphago-mastering-the-ancient-game-of-go-with-machine-learning/>
8. AlphaGo Zero: Revolutionizing AI through Self-Play and Beyond, accessed on May 18, 2025, <https://www.toolify.ai/ai-news/alphago-zero-revolutionizing-ai-through-selfplay-and-beyond-3421281>
9. [2502.10303] Reinforcement Learning in Strategy-Based and Atari Games: A Review of Google DeepMinds Innovations - arXiv, accessed on May 18, 2025, <https://arxiv.org/abs/2502.10303>
10. DeepMind's AlphaGo Zero Becomes Go Champion Without Human Input, accessed on May 18, 2025, <https://futureoflife.org/recent-news/deepminds-alphago-zero-becomes-go-champion-without-human-assistance/>
11. DeepMind's AI agent MuZero could turbocharge YouTube - BBC, accessed on May 18, 2025, <https://www.bbc.com/news/technology-55403473>
12. DAO to HANOI via DeSci: AI Paradigm Shifts from AlphaGo to ..., accessed on May 18, 2025, <https://www.researchgate.net/publication/369707448_DAO_to_HANOI_via_DeSci_AI_Paradigm_Shifts_from_AlphaGo_to_ChatGPT>
13. Alpha Go | AI REV - a boutique AI consulting company, accessed on May 18, 2025, <https://airev.us/alpha-go>
14. AlphaGo - Notes on AI, accessed on May 18, 2025, <https://notesonai.com/alphago>
15. kam.mff.cuni.cz, accessed on May 18, 2025, <https://kam.mff.cuni.cz/~hladik/OS/Slides/Ha-GO-abs-2016.pdf>
16. The computer that mastered Go. Nature video on deepmind's Alpha GO. - Reddit, accessed on May 18, 2025, <https://www.reddit.com/r/MachineLearning/comments/42ymo8/the_computer_that_mastered_go_nature_video_on/>
17. AlphaGo Zero: Starting from scratch - Google DeepMind, accessed on May 18, 2025, <https://deepmind.google/discover/blog/alphago-zero-starting-from-scratch/>
18. AlphaGo Zero - Wikipedia, accessed on May 18, 2025, <https://en.wikipedia.org/wiki/AlphaGo_Zero>
19. What is the "novel reinforcement learning algorithm" in AlphaGo Zero?, accessed on May 18, 2025, <https://datascience.stackexchange.com/questions/23922/what-is-the-novel-reinforcement-learning-algorithm-in-alphago-zero>
20. Why DeepMind AlphaGo Zero is a game changer for AI research - Packt, accessed on May 18, 2025, <https://www.packtpub.com/en-us/learning/how-to-tutorials/deepmind-alphago-zero-game-changer-for-ai-research>
21. augmentingcognition.com, accessed on May 18, 2025, <https://augmentingcognition.com/assets/Silver2017a.pdf>
22. Policy or Value? Loss Function and Playing Strength in AlphaZero-like Self-play - LIACS, accessed on May 18, 2025, <https://liacs.leidenuniv.nl/~plaata1/papers/CoG2019.pdf>
23. Simple Alpha Zero - Surag Nair, accessed on May 18, 2025, <https://suragnair.github.io/posts/alphazero.html>
24. AlphaGo Zero: Discovering new knowledge - YouTube, accessed on May 18, 2025, <https://www.youtube.com/watch?v=WXHFqTvfFSw>
25. AlphaZero - Wikipedia, accessed on May 18, 2025, <https://en.wikipedia.org/wiki/AlphaZero>
26. AlphaZero - Chessprogramming wiki, accessed on May 18, 2025, <https://www.chessprogramming.org/AlphaZero>
27. AlphaZero and MuZero - Google DeepMind, accessed on May 18, 2025, <https://deepmind.google/research/breakthroughs/alphazero-and-muzero/>
28. Alphazero :: Computer-bridge1, accessed on May 18, 2025, <https://www.computerbridge.se/alphazero/>
29. Deepmind's AlphaZero shows unprecedented growth in AI, masters 3 different games, accessed on May 18, 2025, <https://www.packtpub.com/de-no/learning/tech-news/deepminds-alphazero-shows-unprecedented-growth-in-ai-masters-3-different-games>
30. courses.grainger.illinois.edu, accessed on May 18, 2025, <https://courses.grainger.illinois.edu/cs440/fa2021/lectures/games-figs/alphazero.pdf>
31. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm - David Silver, accessed on May 18, 2025, <https://www.davidsilver.uk/wp-content/uploads/2020/03/alphazero.pdf>
32. icml.cc, accessed on May 18, 2025, <https://icml.cc/media/icml-2022/Slides/18016.pdf>
33. werner-duvaud/muzero-general - GitHub, accessed on May 18, 2025, <https://github.com/werner-duvaud/muzero-general>
34. [2411.04580] Interpreting the Learned Model in MuZero Planning - arXiv, accessed on May 18, 2025, <https://arxiv.org/abs/2411.04580>
35. MuZero Intuition - The Waving Cat, accessed on May 18, 2025, <https://thewavingcat.com/2020/12/muzero-intuition/>
36. MuZero - Mastering Atari, Go, Chess and Shogi by Planning with a ..., accessed on May 18, 2025, <https://www.furidamu.org/blog/2020/01/07/muzero--mastering-atari-go-chess-and-shogi-by-planning-with-a-learned-model/>
37. Doubt about MuZero : r/reinforcementlearning - Reddit, accessed on May 18, 2025, <https://www.reddit.com/r/reinforcementlearning/comments/1b0g4gr/doubt_about_muzero/>
38. MuZero beats AlphaZero - Life In 19x19, accessed on May 18, 2025, <https://lifein19x19.com/viewtopic.php?f=18&t=16198&view=next>
39. DeepMind's MuZero is One of the Most Important Deep Learning ..., accessed on May 18, 2025, <https://www.kdnuggets.com/2021/01/deepmind-muzero-important-deep-learning-systems.html>
40. Interpreting the Learned Model in MuZero Planning - ResearchGate, accessed on May 18, 2025, <https://www.researchgate.net/publication/385630316_Interpreting_the_Learned_Model_in_MuZero_Planning>
41. MuZero Explained | Papers With Code, accessed on May 18, 2025, <https://paperswithcode.com/method/muzero>
42. openreview.net, accessed on May 18, 2025, <https://openreview.net/pdf/4f2c0c813d0fbe127329c69b1ba216fbcd95d52c.pdf>
43. ReflectionAI founder Ioannis Antonoglou: From AlphaGo to AGI | Sequoia Capital, accessed on May 18, 2025, <https://www.sequoiacap.com/podcast/training-data-ioannis-antonoglou/>
44. Superhuman artificial intelligence can improve human decision-making by increasing novelty | PNAS, accessed on May 18, 2025, <https://www.pnas.org/doi/10.1073/pnas.2214840120>