智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险

陈小平

引用本文:
陈小平. 人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险[J]. 智能系统学报, 2020, 15(1): 114–120.
CHEN Xiaoping. Criteria of closeness and strong closeness in artificial intelligence——limits, application conditions and ethical risks of existing technologies[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(1): 114–120.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202001001

DOI: 10.11992/tis.202001001

# 人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险

陈小平
（中国科学技术大学 计算机科学与技术学院，安徽 合肥 230026）

**摘　要**：针对现有人工智能技术的两种代表性途径——暴力法和训练法，以及它们结合的一种典型方式，给出了规范化描述，AI 研究中的知识被重新定义为从模型到现实场景的完闭降射，进而提出人工智能的封闭性准则和强封闭性准则。封闭性准则刻画了暴力法和训练法在理论上的能力边界；强封闭性准则刻画了暴力法和训练法在工程中的应用条件。两项准则还为开放性人工智能技术的进一步研究提供了新的概念基础。

**关键词**：人工智能；封闭性；强封闭性；知识；降射；决策论规划；推理；深度学习
**中图分类号**：TP18　　**文献标志码**：A　　**文章编号**：1673−4785(2020)01−0114−07

**中文引用格式**：陈小平. 人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险 [J]. 智能系统学报, 2020, 15(1): 114–120.
**英文引用格式**：CHEN Xiaoping. Criteria of closeness and strong closeness in artificial intelligence——limits, application conditions and ethical risks of existing technologies[J]. CAAI transactions on intelligent systems, 2020, 15(1): 114–120.

# Criteria of closeness and strong closeness in artificial intelligence——limits, application conditions and ethical risks of existing technologies

CHEN Xiaoping
(School of computer science and technology, University of Science and Technology of China, Hefei 230026, China)

**Abstract**: Criteria of closeness and strong closeness in artificial intelligence (AI) are proposed in this paper. The first criterion suggests that knowledge in AI takes conceptual root in a kind of pragmatic correspondence, called consummated grounding, from a model to the scenario that the model is expected to represent. Consummated grounding is critical to advancing both development and explanation of intelligent systems. Under the condition of the second criterion, which aims at real-world applications, existing AI technology surpasses human beings in the same kind of ability, can be successfully applied to realize a lot of projects in current industries, and will not be out of control in itself. The criteria also set up a further conceptual basis for developing AI technology competent to deal with open scenarios.

**Keywords**: artificial intelligence; closeness; strong-closeness; knowledge; grounding; decision-theoretical planning; reasoning; deep learning

在计算机科学中，以图灵可计算性为“可计算”的判别标准，以算法复杂度为实际“计算”的可行性度量，形成了一套完整的计算可解性度量体系，给出“计算”问题在不同抽象或具体层次上可解的标准和条件[1-2]。显然，人工智能也不可缺少类似的可解性度量体系，而且这种需要已十分迫切。

然而，人工智能尚未形成对应于图灵机和 Church-Turing 论题的人工智能理论基础，也没有形成“智能”问题的可解性度量体系，迄今未获得“智能”问题在不同应用场景中的可解性判别标准和可解性条件。为此，本文转换思路，以 60 多年

**收稿日期**：2020−01−02.
**基金项目**：国家自然科学基金项目 (U1613216).
**通信作者**：陈小平. E-mail: xpchen@ustc.edu.cn.

来人工智能研究成果的技术概括为基础，尝试给出现有人工智能技术的能力边界和应用条件。

人工智能研究已形成了至少几千种不同的技术路线，其中很多技术路线可用人工智能的两种经典思维[3]加以概括。这种概括使得对人工智能的认识不再局限于技术路线层面，而上升到“机器思维”的高度。两种经典思维并不代表人工智能的全部，但它们在近期应用中具有关键作用，也是本文构建人工智能可解性度量体系尝试的最可依赖的现有基础，本文将概述人工智能两种经典思维以及它们的集成。在此基础上，本文提出封闭性准则以可刻画现有人工智能技术的能力边界，提出强封闭性准则以刻画现有人工智能技术的应用条件，并重点回答下列问题：现有人工智能技术能解决什么问题、不能解决什么问题？现有人工智能技术能不能大规模产业应用？什么条件下可以应用、什么条件下难以应用？人工智能当前风险如何？最后，简要讨论封闭性准则和强封闭性准则在人工智能基础研究、产业应用和伦理体系建设中的意义与作用。

## 1　基于模型的暴力法

假设 $D$ 是现实世界中的一个应用场景。$D$ 的一个领域模型 $M(D)=<M, Q, g>$是一个三元组，其中 $M$ 是一个数学结构，称为领域模型，由一组参数和它们之间的关系构成；$Q$ 是 $D$ 中的一组待解问题；$g$ 是一个广义映射，称为降射 (grounding)，将 $M$ 的参数和参数间关系对应于 $D$ 的要素和要素间关系，并保持参数间关系 (参数 $p$ 和 $p'$ 在 $M$ 中有关系 $r$ 当且仅当 $g(p)$ 和 $g(p')$ 在 $D$ 中有关系 $g(r)$)，使得 $Q$ 描述的 $D$ 中的现实问题可以在 $M$ 上抽象地求解。由于 $D$ 通常不是一个数学论域，所以一般情况下降射 $g$ 不是数学映射，也无法被形式化表达。$M$ 的参数和关系合称为 $M$ 的元素。

第 1 种人工智能经典思维是基于模型的暴力法，其基本原理是：1) 建立场景 $D$ 的一个精确模型 $M(D)=<M, Q, g>$；2) 构建一个表示 $M$ 的知识库或状态空间 $K^M$，选择一个推理机 reasoner[3-4] 或一种搜索算法[5]searcher，得到扩展模型 $M^*(D)=<M(D), K^M$, reasoner/searcher>，使得在 $K^M$ 上 reasoner 的推理或 searcher 的搜索是计算可行的；3) 对于 $Q$ 中的任何问题 $q$，在 $K^M$ 上用 reasoner 或 searcher 找出 $q$ 的一个正确回答。使用推理机或搜索算法的暴力法分别称为推理法和搜索法。

暴力法的基本前提是：应用场景存在一个精确模型 $M(D)=<M, Q, g>$，其中 $M$ 是良定义的、精确的符号结构。为了使用推理法，需要将精确模型 $M$ 表达成一个知识库 $K^M$；当采用搜索法时，需要将 $M$ 表达成一个状态空间 $K^M$。推理机和搜索算法往往由专业团队长期研发而成，而 $K^M$ 则需由每一个具体应用的开发者手工编写。

以命题逻辑中的推理法为例。在命题逻辑[6]中将 $M$ 表达成一个知识库 $K^M$，相应的推理机 reasoner 一般也是基于命题逻辑的，$Q$ 也要在命题逻辑中表达。对于任何 $q \in Q$，当在命题逻辑中有 $K^M \vdash q$ 时，reasoner 回答 yes；当 $K^M \vdash \neg q$ 时，reasoner 回答 no(其中 $\neg q$ 代表 $q$ 的否定)。因此，reasoner 的开发并非仅仅依靠开发者的直觉，而是以命题逻辑为严格标准，其正确性证明是有理论保证的 (虽然具体证明可能存在各种实际困难)，这种情况在工程方法论中称为可证正确性。

例如，“就餐”场景的有关知识可以人工编写为一个知识库，其中部分知识如表 1 所示，推理机对一些问题的回答如表 2 所示，注意这些回答并不包含在知识库中。

**表 1　一个知识库的例子**

**Table 1　An example of knowledge base**

| 就餐知识的逻辑表达 | 对应的语义解释 |
|---|---|
| $\forall x \forall y$ (dish ($x$) $\wedge$ food ($y$)$\rightarrow$hold ($x$, $y$)) | 餐具可以盛食物 |
| food (rice) | 米饭是食物 |
| food (soup) | 汤是食物 |
| dish (bowl) | 碗是餐具 |

**表 2　一些问答的例子**

**Table 2　Instances of question-answer**

| 问题 | 问题的语义解释 | 回答 |
|---|---|---|
| hold (bowl, rice)? | 碗能盛米饭？ | yes |
| hold (bowl, soup)? | 碗能盛汤？ | yes |
| hold (bowl, $x$)? | 碗能盛什么？ | rice, soup… |

一般地，如果一个逻辑系统具有可靠性，那么该系统中的推理具有保真性[6]。保真性的含义是“结论保持前提的真”，即只要推理的前提在任何一种意义上是“真的”，则推理的结论在相同的意义上也是“真的”。所以，对于任何一个 $M^*(D)=<M(D), K^M$, reasoner/searcher>，如果推理机 reasoner 基于一个具有可靠性的逻辑系统，并且 $K^M$ 是“真的”，则对 $Q$ 中任何问题 $q$ 的回答都是“真的”。这表明，一个具有保真性的推理系统可以应用于任何一个具体场景——不管该场景中“真”的具体含义是什么，只要在该场景中“真”的含义保持一致就可以应用。这就为推理法的普遍应用

奠定了坚实的理论基础。

可证正确性是一种比可解释性强得多的数学性质，而且是迄今为止人类所建立的最强意义上的“可靠性”性质。换言之，在整个科学中没有比保真性更强的通用可靠性机制，工程上也没有比可证正确性更强的可靠性概念。这是暴力法在人工智能三次浪潮中延续不断，并占据第一次和第二次浪潮主流的根本原因。

通常认为暴力法的主要障碍在于知识获取[4,7-8]。一个知识库的“正确性”以及相对于一个应用场景的“充分性”，至今没有形成公认的标准，也没有形成知识库建造的有效技术，致使知识库构建比推理机构建困难得多[4, 7]，暴力法的理论优势——保真性和可证正确性——的效力受到根本性限制。

## 2　基于元模型的训练法

元模型 (meta-model) 是模型的模型。元模型的表达形式可以是形式化的 (如在二阶逻辑中建立的一阶逻辑系统的元模型)，但通常是非形式化的。训练法在不同情况下需要建立不同的元模型，一个应用场景 $D$ 的元模型通常至少包含一组包含标注的“标准数据”集 $T$ 和一套评价准则 $E$，记为 $\mathcal{M}(D)=<T, E>$。评价准则 $E$ 规定了 $D$ 的待解问题集 $Q$ 及求解标准，如求解图像分类问题的一个基本评价指标是分类错误率。

第 2 种人工智能经典思维是基于元模型的训练法，其基本工作原理是：1) 针对应用场景 $D$，设计元模型 $\mathcal{M}(D)$，采集标准数据集 $T$、确定评价准则 $E$；2) 依据 $\mathcal{M}(D)$，选择一种合适的人工神经网络 $m$ 和一个合适的学习算法 $t$，得到扩展的元模型 $M^*(D)=<T, E, m, t>$；3) 依数据拟合原理，以 $T$ 中部分数据为训练数据，用算法 $t$ 训练人工神经网络 $m$ 的连接权重，使得训练后 $m$ 的输出总误差最小。

如果依据 $\mathcal{M}^*(D)=<T, E, m, t>$，训练后 $m$ 达到了 $E$ 规定的全部要求，则称 $\mathcal{M}^*(D)$ 是训练成功的。$\mathcal{M}^*(D)$ 训练成功意味着：依据 $E$ 规定的标准，用训练法成功地解决了 $D$ 中的待解问题。例如，在图像分类任务中，一些经过训练的深层神经网络在给定数据集 $T$ 上的分类错误率已低于人类的错误率，如果评价标准 $E$ 是“在给定数据集 $T$ 上的分类错误率低于人类的错误率”，则这个图像人类任务是训练成功的。

训练法隐含着必须人工完成的大量工作，包括：设计学习目标、决定评价准则、采集数据并标注、选择/设计学习算法、选择测试平台和工具、设计测试方法等。例如在 ImageNet 图像分类比赛中，组织者对大量原始图片中的动物或物品标注一个分类号 (0~999 的整数)，如表 3 所示。其中，将图片分成 1 000 类不是由训练过程自主完成的，而是设计者做出的一项决策。

**表 3　一个图像分类问题**

**Table 3　A sample problem of image classification**

| 原始数据 | 人工标注 |
|---|---|
| 7种鱼的图片 | 0~6 |
| 公鸡、母鸡图片 | 7~8 |
| 26种鸟的图片 | 9~34 |
| ⋮ | ⋮ |
| 卫生纸图片 | 999 |

训练法的理论基础进展状况远远落后于暴力法，不仅没有可证正确性，甚至没有可解释性，这是训练法基础理论研究面临的巨大挑战。

## 3　暴力法与训练法的互补集成

训练法和暴力法都存在明显短板。工程上，训练法的主要短板之一是需要对大量原始数据进行人工标注，暴力法的主要短板是需要人工编写知识库或制定状态空间。一定条件下，暴力法和训练法的结合可同时消除或减弱上述两大短板，AlphaGo Zero[9] 是这种尝试的一个成功案例。“集成智能”已成为未来发展的首要优先方向[10]。

AlphaGo Zero 的暴力法模型是对经典 MDP 模型的修改。一个经典 MDP 模型[11] 的主要元素包括：状态 $s$(代表棋局)、行动 $a$(代表落子)、状态转移函数 $T(s, a, s')$、回报函数 $r$、状态值函数 $V(s)$、行动值函数 $Q(s, a)$ 等。AlphaGo Zero 对这些元素的定义做了修改，核心的改变是将状态值函数 $V(s)$ 的定义从“棋局 $s$ 的期望效用”改为“棋局 $s$ 下的己方平均胜率”，从根本上明确了 AlphaGo Zero 的核心思想，为 AlphaGo Zero 各模块的协调一致构建了统一的基础架构，为 AlphaGo Zero 的巨大成功奠定了坚实的理论基础 (详细说明见文献 [3])。

经上述修改，形成了 AlphaGo Zero 暴力法模型 $M(D)=<M, Q, g>$中的 $M$。$M(D)$ 中的待解问题 $Q$ 是一个博弈策略 $\pi(s)$，其中 $\pi$ 是从任意棋局 $s$ 到落子 $a$ 的射影，$Q$ 的直观含义是：对任意一个棋局 $s$，通过 $M$ 上的推理或搜索，找出 $s$ 上的最佳落子 $a=\pi(s)$。表 4 总结了 AlphaGo Zero 的暴力法模型 $M(D)$ 中 $M$ 和 $Q$ 的主要元素，以及与经典的围棋决策论规划模型的对照。

**表 4　AlphaGo Zero 决策论规划模型 $M(D)$ 及与经典决策论规划模型的对照**

**Table 4　Models of AlphaGo Zero and the standard MDPs**

| 模型元素 | 经典决策论规划模型 | AlphaGo Zero的暴力法模型 |
|---|---|---|
| 回报$r$ | 胜负多少、行动代价等定量值 | $r_T \in \{+1, -1\}$ (表示胜/负) |
| 状态值函数$V(s)$ | 棋局$s$的期望效用(效用的数学期望) | 棋局$s$下己方的平均胜率 |
| 行动值函数$Q(s, a)$ | 棋局$s$下落子$a$的期望效用<br>$Q(s,a)=r(s,a)+\gamma\sum_{s'\in S}T(s,a,s')V(s')$ | 棋局$s$下落子$a$的平均胜率<br>$Q(s,a)=\sum_{s'\|a,s\to s'}V(s')/N(s,a)$ |
| 策略表示$a=\pi(s)$ | 符号表示 | 深层神经网络表示$f_\theta(s)=<P(s,\ -),\ V(s)>$<br>$P(s, -)$是19×19+1个可能落子的概率 |

同时, AlphaGo Zero 还建立了一个训练法的元模型 $\mathcal{M}(D)=<T, E>$, $T$ 的初值设为空集, $E$ 只含一个指标“赢棋”, 因为 AlphaGo Zero 只求获胜, 不考虑赢多少、用时多少等其他指标。AlphaGo Zero 还建立了一个训练法的扩展元模型 $\mathcal{M}^*(D)=<T, E, m, t>$, 其中 $T$ 的数据是通过 AlphaGo Zero 的自博自动产生的, 每一条数据包括一局自博所产生的棋局序列、落子序列和胜负结果, 以胜负结果 (1/−1) 作为标注; $t$ 是强化学习算法; $m$ 是一个最终表示 $\pi(s)$ 的残差网络, 它的输入是任意棋局 $s$, 输出是 19×19+1 个概率值, 分别表示棋局 $s$ 下棋盘上 19×19 个点和 pass 的己方平均胜率。训练好的 $m$ 就是 AlphaGo Zero 的自学结果, 也就是待解问题 $Q$ 的“回答”(即博弈策略 $\pi(s)$)。在训练完成后的对弈实战中, 对任意棋局 $s$, AlphaGo Zero 选择 $m$ 输出的概率最高的点或 pass, 作为自己的最佳落子位置 (即最优博弈决策)。

AlphaGo Zero 的求解过程如下: 1) 构建围棋的暴力法模型 $M(D)$ 和训练法元模型 $\mathcal{M}(D)=<T, E>$。2) 选择蒙特卡洛树搜索作为 searcher, 在 $M(D)$ 上进行 2 900 万局自博, 自动收集每一局自博所产生的棋局序列、落子序列和胜负结果作为 $T$ 的一条数据。3) 设置扩展的元模型 $\mathcal{M}^*(D)=<T, E, m, t>$, 其中 $T$ 更新为第 2 步收集的数据集, $t$ 是强化学习算法, $m$ 是最终表示博弈策略 $\pi(s)$ 的残差网络, 以 $T$ 中数据用 $t$ 训练 $m$, 使 $m$ 输出的总偏差最小[3]。

AlphaGo Zero 带来如下观察: 1) 一定条件下暴力法可以克服训练法的人工标注难点, AlphaGo Zero 利用暴力法的决策论规划模型和蒙特卡洛树搜索, 自动获得了强化学习所需的数据及精确标注。2) 一定条件下训练法可以克服暴力法的知识获取难点, AlphaGo Zero 无需人工编写大量难以形式化的围棋博弈知识, 而是通过强化学习直接获得围棋博弈策略 $\pi(s)$。3) 暴力法和训练法的结合可部分改变训练法缺乏可解释性的缺陷, 得到一种宏观可解释性。例如, AlphaGo Zero 系统的核心构件——建模、自博、搜索、强化学习和围棋博弈决策, 都是围绕平均胜率展开的, 故平均胜率就是 AlphaGo Zero 系统的宏观解释, 解释了该系统的宏观行为原理。欠缺的是微观解释: 为什么 AlphaGo Zero 的残差网络表示的是平均胜率? 有什么保证?

## 4　封闭性

目前在人工智能中可定义两种封闭性——依模型封闭性和依训练封闭性。如果一个应用场景依模型封闭或依训练封闭, 则该场景具有封闭性。两种封闭性分别刻画了暴力法和训练法在理论上的能力边界——不具备封闭性的应用场景, 在理论上无法用暴力法或训练法求解, 至少不存在可解的理论保证。因此, 封闭性是一个极为重要的理论性指标, 故又称为封闭性准则[11]。

### 4.1　依模型封闭

一个应用场景 $D$ 是依模型封闭的, 如果存在一个满足下列全部条件的模型 $M(D)=<M, Q, g>$: 1) 问题确定性: $Q$ 中任何一个问题 $q$ 的回答 $A(q)$ 是唯一确定的; 2) 模型可计算性: 存在扩展模型 $M^*(D)=<M(D), K^M, \text{reasoner/searcher}>$, 使得 reasoner 或 searcher 是图灵可计算的, 并且推理/搜索的结果正确 (以 $A(q)$ 为标准); 3) 降射完闭性: $g$ 是完闭的, 即每一个 $M$ 元素 $e$ 都在 $D$ 中存在唯一、固定的对应物 $g(e)$, 不同的 $e$ 对应于不同的 $g(e)$, 并且对应物集合 $\{g(e)\in D|e \text{ 是 } M \text{ 的一个元素}\}$ 包含场景 $D$ 的所有不可忽略的要素。

“问题确定性”要求的必要性说明如下。在计算机科学中, 称一个函数 $f(x)$ 是图灵可计算的, 首先预设对所有 $x$, $f(x)$ 的值是确定的[1]。然而人工智能面对的很多场景不满足这个预设, 例如在开放领域人机对话中, $Q$ 包含哪些问题、这些问题的“正确”回答是什么, 往往是不确定的, 即使对话系统对所有问题都给出了回答, 也不能确定

回答是否“正确”。因此，仅仅要求对话系统的 reasoner 或 searcher 是图灵可计算的，并不完全符合图灵可计算性的本意。

降射完闭性是人工智能中最难以把握、最易被忽视、最具挑战性的。在简单场景的建模中，完闭性往往被不知不觉地违反，从而导致难以觉察的错误；而在很多复杂场景的建模中，满足完闭性要求通常是极其困难的[4]。导致降射不完闭的 3 种常见情况[3] 如下：1) 对象不确定性——某些对象变体的分类规则难以穷尽地显式表达，导致分类困难，比如即使概念“杯子”在模型中的内涵描述是明确的，其外延和降射却可能无法确定；2) 属性不确定性——现实场景的某些属性是含糊的和场景依赖的，难以穷尽地显式表达，导致这些属性在真实世界中的对应物难以确定；3) 关联不确定性——对象/属性与场景在真实世界中的关联难以确定，也难以穷尽地显式表达。这是暴力法在理论基础研究中遇到的深层挑战。

理论上，如果一个场景是依模型封闭的，则用暴力法是可解的，即存在推理机或搜索算法，对 $Q$ 中每个问题给出正确的回答；反之，一个场景只要不满足 3 个条件中的任何一条，就是非封闭的，该场景用暴力法在理论上是不可解的，或至少没有可解的理论保证。

### 4.2 依训练封闭

首先定义“代表集”。任给训练法的一个扩展的元模型 $\mathcal{M}^*(D)=<T, E, m, t>$，假设场景 $D$ 的全体相关数据的集合为 $T^*$。$T^*$的一个子集 $T^\circ$ 称为 $D$ 的一个代表集，如果 $T^\circ$ 的训练效果不低于 $T^*$的训练效果，即，如果以 $T^*$中数据用 $t$ 训练出的人工神经元网络 $m$ 能达到 $E$ 的全部指标，则以 $T^\circ$ 中数据用 $t$ 训练出的 $m$ 也能达到 $E$ 的全部指标。

例如，假设 $D$ 是一个图像分类任务，$E$ 规定的指标是分类错误率 $\varepsilon$。如果用 $T^*$训练出的人工神经元网络 $m$ 的错误率不高于 $\varepsilon$，那么用 $T^\circ$ 训练出的人工神经元网络 $m$ 的错误率也不高于 $\varepsilon$，则 $T^\circ$ 是 $D$ 的一个代表集。实际应用中，$T^*$通常是得不到的，只能利用它的某个子集，可是并非 $T^*$的任意子集 $T$ 都能够保证训练效果。故本文引入代表集 $T^\circ$。

一个应用场景 $D$ 是依训练封闭的，如果存在 $D$ 的一个元模型 $\mathcal{M}(D)=<T^\circ, E>$，满足下列 2 个条件：1) $T^\circ$ 是 $D$ 的一个代表集，并且是有限确定的，即 $T^\circ$ 是一个有限集，它的每一条数据的内容包括标注都是完全给定的；2) 存在一个扩展的元模型 $\mathcal{M}^*(D)=<T^\circ, E, m, t>$，使得 $\mathcal{M}^*(D)$ 是训练成功的。

直观上，一个场景 $D$ 是依训练封闭的，需要具备一套评价准则 $E$、一个有限确定的代表集 $T^\circ$、一种合适的人工神经元网络 $m$ 和一个合适的学习算法 $t$，使得以 $T^\circ$ 用 $t$ 训练后 $m$ 达到 $E$ 的全部评价指标。其中，学习算法 $t$ 被默认为图灵可计算的。一个场景 $D$ 能否获得满足以上条件的 $E$、$T^\circ$、$m$ 和 $t$，通常没有理论保障，只能依靠训练者的经验和摸索。

理论上，如果一个场景是依训练封闭的，则用训练法可解；反之，如果一个场景不是依训练封闭的，则是不可解的，或至少可解性没有理论保证。例如，如果场景 $D$ 不存在代表集 $T^\circ$，则“扩展的元模型 $\mathcal{M}^*(D)=<T^\circ, E, m, t>$训练成功”是无定义的。

## 5 强封闭性

封闭性准则给出了暴力法和训练法在理论上的能力边界。但是，封闭性准则要求的所有条件都默认为理论上成立，这不符合实际应用的要求，导致满足封闭性准则的场景在工程上仍然不可解。例如，依模型封闭要求存在满足一定条件的扩展模型，其中的“存在”默认为理论上存在，而不是在工程应用中实际地构建出来。

对封闭性准则的另一项重大挑战来自脆弱性。自 20 世纪 80 年代以来，脆弱性已成为现有人工智能技术实际应用的主要瓶颈，训练法和暴力法都深受其害。脆弱性的主要表现是：如果智能系统的输入不在知识库或训练好的人工神经网络的有效范围内，系统可产生错误的输出。实际应用中无处不在的感知噪声是脆弱性的一个主要原因。例如，在文献 [12] 报告的测试中，先用一个著名商用机器学习系统训练出一个深层神经网络，该网络可以很低的误识别率从照片中识别各种枪支。然后，人为修改这些照片上的少量像素(代表感知噪声)，而这些修改对人眼识别没有任何影响，可是训练好的深层神经网络对于被修改照片的误识别率却大幅升高，而且会发生离奇的错误。2013 年以来，针对深度学习已发现大量类似的例子。

上述困难目前在理论上无解，但一定条件下是工程上可解的。本文将这些条件概括为强封闭性准则，在符合该准则的工程项目中可应用暴力法、训练法或它们的集成。

一个场景 $D$ 在一个工程项目 $P$ 中具有强封闭性，如果满足下列所有条件：1) 场景 $D$ 具有封闭

性；2) 场景 $D$ 具有失误非致命性，即应用于场景 $D$ 的智能系统的失误不产生致命的后果；3) 基础条件成熟性，即封闭性包含的所有要求在项目 $P$ 中都得到实际满足。

基础条件成熟性要求，暴力法需要的问题确定性、模型可计算性 (包括推理机/搜索算法存在性)、降射完闭性，训练法需要的代表集存在性、元模型存在性、扩展元模型存在性及训练成功等条件，都在工程项目中得到实际满足。因此，强封闭性准则是与具体工程项目相关的，工程团队的实力，工程的工期、投入和其他资源的不同，都可能影响一个工程项目是否符合强封闭性准则。

即使一个工程项目完全满足基础条件成熟性，由于脆弱性的困扰，仍无法保证智能系统不出现失误，包括致命性失误。为此，强封闭性准则引入了失误非致命性要求。失误非致命性和基础条件成熟性往往需要通过场景封闭化才能够满足。目前主要有两种封闭化手段：场景裁剪和场景改造，二者普遍适用于人工智能在信息产业和实体经济行业中的应用。

场景裁剪的原理是：以智能系统的可靠性、安全性为目标，对应用场景进行取舍，排除可能导致致命性失误或违反基础条件成熟性要求的情况。例如，在训练法中，为了规避无法获得代表集的难点，可将应用场景限制在环境变化可忽略或可控的范围内，在这种环境中可以获得质量足够高的训练数据集作为代表集。

场景改造的原理是：以环境可控性为目标，通过对应用场景的改造或部分改造，使之封闭化、准封闭化或局部封闭化。封闭化场景完全符合强封闭性准则；准封闭化场景基本满足基础条件成熟性，同时满足失误非致命性；局部封闭化是在场景的某些局部实现封闭化。场景改造在汽车制造业自动化中取得巨大成功，目前正在快速扩展到其他制造业行业，并且智能化程度不断提升，对农业、服务业的很多部门也是适用的。

对一部分场景 (如开放领域人机对话) 而言，只要符合失误非致命性，即使另外两项要求不完全满足，也可能被接受。关于强封闭性准则的通俗解释见文献 [12]。

## 6　讨论

经历了三次浪潮、仍受封闭性限制的人工智能，当前面临的最大疑问是能否找到大规模应用的可行途径。本文给出了一个回答：符合强封闭性准则的工程项目可成功地应用现有人工智能技术，不符合的不能。现实中，满足强封闭性准则的行业部门大量存在，尤其在制造业、智慧农业等行业。然而，目前大部分人工智能工程项目并不符合强封闭性准则，由此带来的困难和困扰正在呈现出来。

根据封闭性准则，暴力法是基于知识的，而知识本质上是完毕降射，完闭性隐含着对场景中存在的各种“主体”的行为效果的充分把握。因此，知识实际上包含着语用，而人工智能的主流观点将知识理解为单纯的语义，极少考虑语用。文献 [13-15] 建立了机器人概念模型与行动模型之间的语用关联，并为语用的有效表达引入了多型知识。近年来随着可解释性引起关注，机器人概念模型与行动模型的关联得到了更多研究[16]。事实上，语用观点更恰当地反映了“智能问题”与“计算问题”的本质区别：“计算”主要涉及算法及计算资源，而“智能”则广泛涉及对现实世界的把握和交互[5, 14, 17-18]。

近年来，降射引起了人工智能研究者的关注[15, 18-19]，然而暴力法尚未形成降射完闭性的成熟理论和技术。训练法的一个基本出发点是从带标注数据“提取知识”，从而绕过这一核心难点。该努力在封闭性范围内取得了里程碑式进展，在非封闭性条件下未取得预期成功，两种情况下均带来不可解释的新难题。可见降射完闭性是无法回避的一项基础挑战。

封闭性准则给出了暴力法和训练法理论上的能力边界，即在非封闭场景中暴力法和训练法的应用没有成功保证。然而对很多产业部门特别是生活场景而言，封闭化往往是不适用的，发展开放性场景中的人工智能技术是一项长期追求，国内外学者进行了大量尝试和探索[3-4, 14-15, 17, 19-21]。在前期工作的基础上，本文进一步给出了开放性的一种理论上更系统的解释：不满足封闭性准则的是理论上的开放性；不满足强封闭性准则的是工程上的开放性，从而为探索开放性人工智能提供了新的参考。

在强封闭性准则范围内，现有人工智能技术的实际应用通常需要经过由人完成的场景裁剪或场景改造，所以人工智能不可能脱离人类而独立发展，不存在技术失控的风险。另一方面，对人工智能的技术误判已成为当前人工智能发展的首要障碍，将带来应用受阻、管理失误、技术误用等后果和风险。封闭性和强封闭性准则为消除技术误判提供了依据，有助于建立符合技术真实性和社会发展需求的人工智能伦理体系[21]。

## 参考文献：


[1] HOPCROFT J E, ULLMAN J D. Formal languages and their relation to automata[M]. Boston, USA: Addison-Wesley, 1969.

[2] KNUTH D E. The art of computer programming: volume 1: fundamental algorithms[M]. 3rd ed. Redwood City, USA: Addison-Wesley Professional, 1997.

[3] 陈小平. 人工智能的历史进步、目标定位和思维演化 [J]. 开放时代, 2018(6): 31–48.
CHEN Xiaoping. Artificial intelligence: advancement, goals and change of thinking mode[J]. Open times, 2018(6): 31–48.

[4] DAVIS E. The naive physics perplex[J]. AI magazine, 1998, 19(4): 51–79.

[5] NILSSON N J. Artificial intelligence: a new synthesis[M]. 北京: 机械工业出版社, 1999.

[6] 汪芳庭. 数理逻辑 [M]. 2 版. 合肥: 中国科学技术大学出版社, 2010.

[7] LENAT D B, GUHA R V. Building large knowledge-based systems: representation and inference in the cyc project[M]. Reading: Addison-Wesley, 1990.

[8] DAVIS R, SHROBE H, SZOLOVITS P. What is a knowledge representation?[J]. AI magazine, 1993, 14(1): 17–33.

[9] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354–359.

[10] GIL Y, SELMAN B. A 20-year community roadmap for artificial intelligence research in the US[R]. Computing Community Consortium, Association for the Advancement of Artificial Intelligence, 2019. Washington, D.C., USA.

[11] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial intelligence, 1998, 101(1/2): 99–134.

[12] 陈小平. 封闭性场景: 人工智能的产业化路径 [J]. 文化纵横, 2020(2): 34–42.
Xiaoping Chen. Closed contexts: A feasible approach to AI industrialization[J]. Beijing Cultural Review, 2020(2): 34–42.

[13] ILYAS A, ENGSTROM L, ATHALYE A, et al. Query-efficient black-box adversarial examples (superceded)[J]. arxiv: 1712.07113.

[14] CHEN Xiaoping, JI Jianmin, JIANG Jiehui, et al. Developing high-level cognitive functions for service robots[C]// Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. Toronto, Canada, 2010.

[15] CHEN Xiaoping, XIE Jiongkun, JI Jianmin, et al. Toward open knowledge enabling for human-robot interaction[J]. Journal of human-robot interaction, 2012, 1(2): 100–117.

[16] EDMONDS M, GAO Feng, LIU Hangxin, et al. A tale of two explanations: enhancing human trust by explaining robot behavior[J]. Science robotics, DOI: 10.1126/scirobotics.aay4663

[17] TURING A. Intelligent machinery (manuscript)[J]. The turing digital archive, 1948.

[18] TELLEX S, KOLLAR T, DICKERSON S, et al. understanding natural language commands for robotic navigation and mobile manipulation[C]//Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011.

[19] KOLLAR T, SAMADI M, VELOSO M. Enabling robots to find and fetch objects by querying the web[C]//Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. Valencia, Spain, 2012.

[20] CHEN Xiaoping, JI Jianmin, SUI Zhiqiang, et al. Handling open knowledge for service robots[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013.

[21] 陈小平. 人工智能伦理体系: 基础架构与关键问题 [J]. 智能系统学报, 2019, 14(4): 605–610.
CHEN Xiaoping. Ethical system of artificial intelligence: infrastructure and key issues[J]. CAAI transactions on intelligent systems, 2019, 14(4): 605–610.


## 作者简介：

陈小平，教授，中国人工智能学会人工智能伦理道德专委会主任，主要研究方向为人工智能理论基础和智能机器人关键技术。提出基于“开放知识”的机器人智能技术路线，并在“可佳”和“佳佳”智能机器人系统中进行了持续性研究和工程实现。团队自主研发的“可佳”机器人 2015 年获国际服务机器人精确测试第一名，2014 年获国际服务机器人标准测试第一名，2013 年获第 23 届世界人工智能联合大会最佳自主机器人奖和通用机器人技能奖。2005 年以来团队在机器人世界杯上先后获得 12 项世界冠军。多次获得国际学术会议最佳论文奖。获 2010 年度中科大“杰出研究”校长奖。