

人工智能研究需要更高标准

John McCarthy

本文是John McCarthy在1984年AAAI会议上的主题发言。McCarthy提出需要确立AI研究的衡量准则，以推动AI基础研究。这些准则主要包括三个方面：普遍性，容忍性（如框架问题）与认识合适性。它们不仅适合于理论AI研究，而且也是实验AI研究的准则。从更高标准看，目前AI许多领域缺乏实质性进展。值得一提的是，McCarthy批评了图灵测验作为AI的科学准则，他指出，衡量AI程序是否达到人的智能程度的一般准则是看它能否体现人的常识行为，这正是当前AI研究的关键问题。我们相信McCarthy的意见对国内AI研究者有所裨益。

——译者林作铨

任何一门科学的现代发展水平少不了衡量研究的准则。像科学的每一其它方面一样，准则方面也必须加以发展。衡量AI研究的准则尚不清楚。如果我们有衡量AI研究结果的更高标准，那么这一领域的发展就会更快。

我们应当克服不负责任的情况，一篇论文通常报告了一个先前没有做过的计算机程序，但是，这一工作是否推动科学的发展，或其他研究者能否从这一工作中得到帮助，可能并不清楚。

差距、认真对待、改革开放；内外结合必须以内为主、为内服务。士气与信心至为重要，找出差距存在的本质原因，埋头苦干，终必有成。但是政策必须落实，政策不落实，就无法鼓舞士气，士气不振，必然要打败仗。在诸多政策中“按劳分配”尤为重要。软件是智力产品，软件开发是智力密集的活动，软件人员智力高度集中，宜相应制订具体规定，否则高水平的软件人员外流，恶果难以设想。

2. 将于1994年召开的第十六届国际软件工程会议，海外同行至为关注。一些有影响的专家建议争取能在我国召开，并愿意加以

一些人从道义上指责其他人愚弄了出版物与读者，可是没有理由认为AI研究者比其他科学家缺乏进取。我没听说过AI研究的整体水平比其它领域差的说法，我担心的是AI缺少基础研究，而导致这种现象的理由之一是判断一个研究工作是否取得实质性进步是困难的。

研究水平的衡量准则似应基于AI研究所期望的进展，我还不能列出一套完整的衡量准则，下面是一些意见。

1. 假设一个研究是用计算机解决一个

促进。这表明炎黄子孙的爱国热忱。但是，要有一个有说服力的“建议书”，其中须列出切实可行的计划。至为重要的则是经济保证与组织保证。记得1985年在伦敦会议上曾讨论第十届会议的东道国的问题，当时也曾有可能争取到在我国召开，但因准备不足，终于让新加坡争到。明年五月在澳大利亚会议上就要讨论决定，时间很急，请领导裁定。如有需要，我们愿意出力促成。

此次会议内容丰富，关于会后访问情况，将另文汇报。上述各点，不当之处，请批评指正。

先前尚未解决的问题。

假设这一方法缺乏适合解决某一类问题的理论论证，只是有一个能解决一定的样本问题的可执行程序，以及对这一程序的一些解释。

这种研究的合适性是难以说明的，读者不能轻易地确信这一程序不是围绕特定问题发展起来的，因此，其他研究者不能从这一工作中学习多少东西。当然，对于当前的研究现状一个太难的论题有时只好先做出来。

2. 一个较好的结果是描述以前未能说明的智能机制，并证明了对某一类问题是充分或必要的。

一个例子是博弈中的 α - β 启发式方法，人们使用这种方法，但第一个下棋程序的作者并未能证明，启发式方法虽不是下棋程序的全部，但却是必要的，因为棋步的增长是很快的。

3. 实验工作应是可重复的。

实验科学，如物理学与生物学，通常重复以前的实验以验证一定的现象或实验中已确定的常数值。审稿人希望一篇实验工作的论文包含足够的细节以确信同一工作可以重做出来。

实验AI研究在博士论文中最常见，研究者首先描述欲解决的一般问题，然后描述所给出的程序以及对样本问题所获得的相应结果，而这一程序或方法能够解决哪一类问题经常并没有加以说明。这样，读者可以随便怀疑这个程序能否解决论文中特定的问题，甚至作者也不清楚是否还能解别的问题。

如果我们需要可测试或可重复的AI工作，那么论文作者与审稿人必须给出程序的普遍性说明。如果语言，硬件与通讯设施允许的话，审稿人应能够重复做出这种程序的结果。此外，这种方法应描述得足够清楚，便于他人可以直接应用并检测其合适性。

正如其它实验科学那样，AI中重复他人试验一般是正常的，可以说还比较容易做

到，因为AI或多或少有标准的硬件和程序语言可用。也许这对开始学习AI的研究生或AI外的人进入这一领域都有好处，他们可能高兴发现一个反例是某个程序所设想解决却不能解决的。

说明一个研究工作的普遍性在许多方面是困难的，研究者最好在完成了能解决例题的程序后通过理解程序行为来分析它的普遍性，研究者必须做出这种努力，否则会有某个聪明的学生替他找到反例。

4. 我们也需要形式化的准则。

基于逻辑的AI方法要求使用逻辑语言表示常识世界的一般事实，并要求说明推理方法（包括非单调推理）能确定一个机器人如何做出给定的目标（用句子表达），能确定一般事实与特殊事实。判断这些事实的形式化是否成功的主要准则是普遍性与认识合适性。普遍性部分地是语言的性质，就一阶语言而论，普遍性意指谓词与函项符号集。当语言需要扩充时这些符号不必加以修改，普遍性部分地也是公理集的性质，扩充时这些公理不必改变。近来，非单调推理形式化的发展清楚地说明了这些道理。

提出一种逻辑形式化的论文作者应尽可能指出所具有的普遍性，审稿与评论将验证是否达到了这一点。

认识合适性指的是表达一个人或机器人的信息系统可能知道和必须知道的事实的能力。

5. 衡量搜索方法的准则有比其它领域较好的标准。

据我在五十年代后期到六十年代早期对下棋程序的经验，可以很好地分析如 α - β 等各种搜索方法能减少搜索时空。

6. 另一方面，衡量自然语言理解程序的准则是最糟糕的。

这里，人们经常不相信他人主张的普遍性。

对这个领域我提出以下两个挑战性问题：首先，我提供某个故事（按字母顺序的）

数据库论战卷土重来

Christopher M. Stone和David Hentchel

摘要 自“面向对象的数据库系统宣言”和“第三代数据库系统宣言”相继发表之后，当今数据库论战卷土重来。本文作者根据这两个宣言，详细讨论了这两大阵营在理论、观点和技术上的异同，澄清了一些错误看法，并分析了相应的应用领域和产品情况。作者认为论战将以OODB的胜利而告终。

当七十年代伟大的数据库论战偃旗息鼓时，Cood关于关系数据模型的十二条规则战胜了层次和网状模型。然而，在八十年代，一批专门研究面向对象技术的著名学者以著文“面向对象的数据库系统宣言”〔中译文见本刊1990年第三期〕又提出了新的挑战。关系数据库阵营很快响应，著文“第三代数据库系统宣言”〔中译文见本刊1991年第1期〕。从此，数据库设计者和用户的智力投资与购买

预算都要担风险。

或许你领略不到专家们在理论上的唇枪舌剑，但你肯定会对论战的结果感兴趣。论战的焦点是，在未来十年中最适合用户需求的数据存贮类型问题。

关系的行动准则

几位数据库名流（Michael Stonebreaker、Larry Rows、David Beech、Bruce Lindsay等）所著的“第三代数据库系统宣

词汇表和一组有关问题。计算语言系统研究者建立他的系统来理解这个故事并且给出有关问题的回答，当他建立这个系统后，我就进一步提供这个故事与问题，他能用自然语言表达这些问题或翻译成该系统的合适输入。应事先描述该系统的一些局限，然后我们看一看该系统可以成功地回答什么问题，而系统建立者理解这些局限的程度如何。

第二个问题是如何建立这样一个系统，它能从数据库中获得信息并能用自然语言与用户进行交互。再事先给定词汇表，系统研究者调整系统去适应，然后测试系统能否与数据库联接并回答问题。何如，对一个人口普查数据库能否写一个程序给用户确定有关的人口分布信息。

我想这两个问题都是相当困难的，什么时候某个研究小组能很好解决它们尚不得知，或许，这是AAAI奖或其它什么奖的好

题目。

然而，这样的挑战性问题并不能取代衡量自然语言理解研究的准则。

7. 图灵测验曾是一个AI的挑战性问题，但不是一个AI研究的科学准则。

适当修改的图灵测验只是确信AI达到人的智能程度的一个充分条件。然而，我们需要朴素点的主张，说明一个特殊的智能机制是解决某一类问题的充分条件或必要条件。

图灵测验即使作为一个充分条件也需要进行修正。模拟人的能力必须经受一个知道程序原理的人的挑战。此外，我们正处在一种让人们认为是魔术师的局面，我们不能忽视这种危险。何况通过对话就想出某种技巧即是机器（AI）是不合适的。

〔林作铨译自《The AI Magazine》
5(1984) pp.282—285.〕