中国教育财政

怀仁怀朴　唯真唯实

北京大学中国教育财政科学研究所

2018 年第 7-2 期（总第 157 期）　　2018 年 6 月 19 日

人工智能在教育测评领域的应用与研究现状

——教育与人工智能系列谈

黄晓婷*

近年来，人工智能在社会生活的各个领域都得到了越来越广泛的应用，如零售行业中分析消费者消费习惯的商业智能、汽车制造中的自动驾驶等。在教育领域，在线教育在过去十多年里飞速发展，积累了大量的数据，为人工智能的研究奠定了数据基础，也对人工智能的应用提出了新的需求。

一、人工智能在教育领域的主要应用

目前，人工智能在教育领域的应用主要包括四类：

第一类是“行为探测”，如考场的作弊监控系统。类似的应用还有前不久新闻里报道的“魔镜系统”，即通过人脸识别，实时探测学生是否在认真听讲。不过，是否应该在课堂教学中运用这样的系统还存在很大争议。

第二类应用被称为“预测模型”，如通过学生学习过程中的行为数据，预测

* 黄晓婷，北京大学中国教育财政科学研究所副研究员。

学生是否有高辍学风险，或者预测学生成绩是否及格等。已有的研究主要集中在MOOC 领域。学者们使用学生上线时间、观看视频时间、次数、参与讨论情况、作业提交情况等数据，预测学生是否能完成某一课程，从而使教师能及早为有困难的学生提供帮助，提高 MOOC 的效率。

第三类应用为“学习模型”，如在线的自适应学习系统，即根据学生兴趣、学习能力、知识掌握情况等因素，为学生提供适宜的学习内容。有一些研究试图为学生提供符合其认知模式的学习内容，如为对图像敏感的学生提供以视觉刺激为主的学习资料，但目前研究者们还没有发展出非常成熟的应用。

第四类应用“智能测评”与“学习模型”紧密相关。在自适应学习中，系统需要首先对学生的能力、知识掌握情况等进行测评。智能测评旨在以传统测评无法比拟的效率，完成对学生的测评和诊断任务。

二、人工智能在教育测评中的应用

智能测评包括人工智能在传统测试的各个环节中的应用。教育测评的过程本质上是把某种潜在特质（看不见、摸不着又确实存在的能力、素养或心理特质）用一种科学的方法进行量化，用数值来表示被试在该项特质上的发展水平。传统的测评主要有三个环节：命题、答题和评分。人工智能在这三个环节中的应用即为机器命题、机器答题和自动评分。

1．机器命题

传统命题是由学科专家或专业的命题人员，根据考试的目的，设计试题的过程。命题质量是决定整个测评质量的关键因素，整个试卷在内容上应该是所有需要考评的内容的代表性抽样。试卷难度应当满足测试目的：选拔性考试通常偏难，而达标考核的难度则依据相应标准来确定。

在线学习系统和计算机自适应考试的发展，大大增加了对试题数量的需求。一次传统的纸笔考试可能只需要 50 题左右，但在自适应考试中，需要给每个考生不同的试题，所需的题目数量就成倍增加。同时，自适应考试和在线学习系统中测试的频次往往较高，因此也需要更多的试题。传统的命题成本较高，耗费时间较久，且存在一定的错误率，而机器命题能大幅节约命题成本，提高命题效率。此外，由于机器命题没有泄露试题的风险，提高了考试安全性。因此，机器命题

在过去十多年里得到了较快的发展。

机器命题有两种主要的模式：强理论模型和弱理论模型。所谓强理论模型，是指在比较扎实的认知理论基础上进行命题。比如部分数学题，解题所需要的能力可以分解为问题提炼、数学表达、运算执行等几个部分。通过分析一组类似试题的考生作答数据，测量学专家们可以较为精确地计算出每个步骤的难度以及这个步骤在整个题目中的权重。随后，计算机自动替换题目中的一个或几个元素，生成新题。这样的新题可以在“母题”的基础上进行较多的变化，新的难度也在很大程度上可控。

不过，教育领域的大部分考试都缺少对应的认知理论支撑。因此，机器命题更多使用弱理论模型。具体过程大致如下：命题专家先找出性能好的题目作为母题，再对题目进行非常详细的分析，构成多层次的题目模型，即把题目分解成背景、内容、问题、辅助信息与选项等部分。接下来，专家再确定可以替换的部分。计算机先分析可替换部分的文本难度、问题的难度，再从语料库和数据库中找到合适的内容，进行替换，形成新题。这类新题和母题的相似度很高，难度也基本保持不变。

数学和英语是机器命题应用较多的学科，特别是英语的语法和阅读理解题，已经有一些商业软件可以完成命题。例如，“Item Distiller”软件主要被用来命以单句为主的语法题，“EAQC（enhanced automatic question creator）”软件则多用于命阅读理解题。

尽管机器命题能节约成本，提高效率，但也存在一定的局限。首先，命题过程仍然离不开命题专家对母题的选择和分析。其次，机器在设计干扰项时比较死板，只会依据母题的模版生成干扰项，而不会根据题目的特点重新设计。第三，由于开放性问题（如简答题等）的标准答案设计需要另一套设计模型，机器命题目前也较少被用于此类问题。最后，机器命题十分依赖语料库。英语的语料库发展比较快，计算语言学的研究已经完成了对词的难度、词和词之间的距离等的量化，为机器命题奠定了良好的基础。而对其他没有成熟语料库的语言来说，好的

评分员进行打分的开放性问题，如口语考试、简答题、作文题等。评分员打分耗时耗力，机器自动评分可以节约时间和成本，大大提高效率。

自动评分一般包括三个步骤。首先，要把语言或手写的文字转化为电脑可以读取、分析的文本。这一步依赖自然语言处理系统，目前中文也有一些软件可以便捷地完成处理。

第二步，分析文本。常用的分析方法有两种，一种被称为“隐含语义分析”，另一种则是“人工神经网络”。所谓隐含语义分析，是指把被试的回答转换成数字矩阵，计算与标准答案矩阵之间的距离。这种方法多用于简答题。对于较长的回答，如作文，则更多使用人工神经网络。人工神经网络简单说来就是找出本文

大力加快题库建设，但由于保密问题，很难实现在高考这样的高利害考试中使用试测过的试题。机器答题也可以大大降低泄露试题的风险。机器答题的复杂程度更高，目前还没有成熟的、商业化的应用。我国的科大讯飞正在积极研发，日本、欧美也有一些团队在进行研究。

三、人工智能与教育测评的未来研究方向

人工智能在命题、答题和评分中的研究和应用都在不断推进过程中。但不少研究者认为，目前的这些应用没有改变测评的基本内容和形式，只在一定程度上降低了成本、提高了效率。在线学习平台已经积累的数据，应该能够支撑研究者们进行更多的探索，突破原有的测评方式，例如应用学习过程中的行为数据完成测试等。研究者们开创了一个新的领域——“分析测量学”，即通过大数据分析而非传统的考试，对学生进行测评。

墨尔本大学教育学院的研究团队已经进行了初步的探索。他们通过分析学生在一项游戏化学习过程中的 1600 多个行为数据，对学生的合作问题解决能力、批判性思维能力、创新领导力等几项核心素养进行评估。分析测量学仍然遵循测量学的基本逻辑：首先要建立理论框架；随后在学科和认知理论的基础上，进行新型“命题”，即通过数据挖掘找到高相关的信息，同时通过传统命题的思路赋予这些数据实践意义；随后再通过理论与数据结合的方式，对不同的行为进行评分；最后运用测量学模型估算被试的能力。这种“分析测量”将改变测试的场景、命题和评分方式，给测量领域带来更具深远意义的变革。

人工智能在高效实现个性化学习方面有着无可比拟的优势，未来在教育领域的应用必将更为广泛。但在我们热情迎接人工智能时代的同时，研究者和实践者们仍需保持谨慎。人类认知的拼图还远没有拼完整，因此我们要理智地对待根据已有大数据得出的结论，防止推论过度泛化。此外，如何保护学生、教师和学校的隐私和秘密，合理使用数据，也是急需我们思考和解决的问题。

上期回顾

2018 年第 7-1 期（总第 157 期）

人工智能浅析——教育与人工智能系列谈

摘要：为帮助缺乏相关背景知识的读者更系统地了解人工智能，本文从发展历史、现代研究体系、与大数据的关系及专家观点分享几个部分对人工智能进行了介绍。

《中国教育财政》由北京大学中国教育财政科学研究所主办；旨在反映本所最新的学术科研活动；相关内容仅体现作者本人观点，并不必然代表本所的立场。

文章内容仅供参考，如需转载须事先征得本研究所同意。

本期印发：1900 份

下载网址：http://ciefr.pku.edu.cn

主办单位：北京大学中国教育财政科学研究所

责任编辑：毕建宏

电　　话：010-6275-9700

地　　址：北京市海淀区颐和园路 5 号

北京大学教育学院楼 413 室（100871）

电子信箱：newspaper@ciefr.pku.edu.cn

传　　真：010-6275-6183

微信公众号：中国教育财政