国际视野下人工智能教育测评：应用、风险及规避

徐湘荷

（聊城大学 教育科学学院，山东 聊城 252000）

摘　要：人工智能教育测评是“数智化”时代教育测评的发展趋势，包括自动化评分系统、计算机自适应测试、基于过程数据的形成性测评和预测以及测评诚信保障机制。相对于传统的人工测评，人工智能教育测评的优势非常明显。但与此同时也面临着诸多的风险，具体来说，就是人工智能教育测评本身的风险、使用人工智能教育测评的风险以及二者共同面临的风险。为了规避这些风险，首先开发和使用人工智能教育测评需要评估它对学生的益处和风险；其次，在测评的任何环节，都应该保护学生的隐私权等各种权利。当然，这些要求实现前提是人工智能教育测评必须以人类为中心，以尊重学生的主体性和自主性为宗旨。一旦出现对学生的侵权和伤害，相关人员能够被问责并有及时的补救措施。

关键词：人工智能教育测评；风险；规避；国际视野

中图分类号：G420　　**文献标识码：**A　　**文章编号：**1672-1217（2025）04-0139-09

收稿日期：2025-05-09

基金项目：聊城大学教学改革项目（311162346）：《强师计划》背景下高等学校公共基础课《教育学》教学改革研究与实践。

作者简介：徐湘荷（1971-），女，山东冠县人，聊城大学教育科学学院副教授，教育学博士。

DOI:10.16284/j.cnki.cn37-1401/c.2025.04.012

科技发展和科技风险总是相伴相随。近年来，人工智能教育测评的发展突飞猛进，它不仅对学生提交的作业、考试文本或图像、音频等自动批改和评分，还能自动生成测试项目，并提供个性化的反馈。人工智能教育测评不仅为教师“精准教”和学生“个性学”提供了帮助，而且提升了教育境界和品质，甚至有改变教育图景之势。但与此同时，人工智能教育测评也面临诸多的风险。本来，人工智能教育风险一直都是学术界、企业界、媒体界和政策立法争论的热点话题。但具体到人工智能教育测评的风险和规避，目前国内外尚未达成共识。本文通过梳理和研究国际社会人工智能教育测评应用，分析其优势、提示其风险，并提供相应的风险规避框架，希冀能为人工智能教育测评应用提供一些指导和帮助。

一、人工智能教育测评：“数智化”时代教育测评的发展趋势

当今社会是向“数智化”时代转型的时期，在这一时期，以数据和数字为基础的智能化技术深度融入各个行业和领域，推动了社会经济、文化、生活方式等方面发生了深刻变革，教育领域也不例外。人工智能教育测评是“数智化”时代教育测评的发展趋势。因为人工智能教育测评不仅是人工智能在教育场景的典型应用，也是人工智能教育场景应用的基础，像智能学习系统、智能辅导系统、虚拟学习助手等都需要根据学习者不同的学习偏好和学习进程，自动化预测和评估学习效果，并智能调整学习内容、学习节奏和学习目标。本文的人工智能教育测评是利用人工智能强大的数据分析和处理能力，通过算法收集、分析大量的教育教学相关数据，如学生表现、作业完成速度和质量、项目完成过程和结果等，对学生的学习过程和结果进行判断和评估。

（一）依据其功能，目前的人工智能教育测评主要类型有：

1. 自动化评分系统。自动化评分系统常见应用之一是根据预设的标准或模型，利用机器学习、自然语言处理等人工智能技术对文本、图像、音频等多种类型的数据进行分析，并给出相应的评分结果。我国中小学“作业帮”“科大讯飞”作业自动批改系统；国际社会批阅英语作文常用的“交上吧”（Turnitin）、“电子评分系统”（E-Rater）都属于这一类。自动化评分系统的另一个应用是根据已有的数据和相应的算法，来

测算学生的成绩。一些学校利用这种工具来招生，早在2013年德克萨斯大学奥斯汀分校计算机科学系开始使用“评分”（Grade）系统来遴选博士资格候选人。2020年春夏英国A-level考试（General Certificate of Education A-level，简称A-level）及国际文凭项目（International Baccalaureate Diploma Programme）的毕业考（这两种考试都类似于中国高考），都因新冠疫情求助于自动化评分系统来评定学生成绩。

2. 计算机自适应测试。传统人工测试使用固定的测试题目评定学生成绩；而计算机自适应测试是利用机器学习，根据考生对测试题目的回答情况动态调整后续的测试项目，并能即时自动评分。起初，计算机自适应测试通常用于利害攸关的总结性评估，如美国教育测试服务中心（ETS）提供的经企管理类研究生入学考试（GMAT）、北美地区研究生申请考试（GRE）等。现在，随着生成性人工智能的兴起，计算机自适应测试可以根据教学任务、学生水平自动生成高质量的测试题目，越来越多的形成性测评使用这种形式发布随堂练习、单元测验等。

3. 基于过程数据的形成性测评和预测系统。在人工智能教学情境中，当学生与计算机互动时，软件可以捕捉到学生所有的活动痕迹并将其存储于数字日志文件。这些活动痕迹包括学生查看各个项目的次数、项目之间移动路径、项目完成过程、完成时间和结果以及整个学习过程中学生活动的全面点击流。依据这些数据可以生成完整、连续的学生参与图和形成性反馈报告。此外，一些学校常用的学业预警系统也属于这一类。人工智能通过分析学生的过程数据如出勤情况、课堂参与度、课堂表现和课堂作业完成的速度和质量等，识别出可能面临学业挫折的学生，从而使得教师和学校能够对这些学生进行针对性的干预。

4. 为保护测评诚信的自动检测系统或者自动监考系统。为了保护教育测评的诚信，绝大部分自动化评分系统都自带检测功能，能检测到所提交的作业或测试是否有学术不端。另外，独立的论文检测系统如“中国知网”能检测到提交论文是否有学术剽窃或抄袭。至于监考系统，传统上利害攸关考试通常在有教师实体监考的场地进行；而2020年新冠疫情加速了人们对远程监考的接受。在线远程监考系统通过网络摄像头对考生周围环境和考生本人的行为、面部表情、眼神等方面的信息进行监控、检查、追踪和记录，以判断学生是否作弊。

5. 互动、体验式人工智能教育测评情境。目前虚拟现实和增强现实工具使用沉浸式技术为学生提供互动、体验式测评情境。在这一情境中，学生通过参与活动，体验解决现实问题的真实情境。

（二）相对于传统的人工测评，人工智能教育测评有以下优势：

1. 教育测评的工作量从繁重到易行。传统人工测评工作可能非常繁重，并且还有出错的风险。因为人工测评需要精心设计测评项目、参考答案或评分标准，随后人工评分以及越来越复杂的成绩统计分析。而人工智能教育测评可以部分或完全自动化测评某些环节，甚至可以智能化生成教育测评项目，并能即时评分和反馈。这样不仅使得教育测评更加方便易行，而且也能为师生提供即时化反馈。心理学的研究表明，学习过程中的即时化反馈有助于及时发现并纠正学生学习中的错误，提高学习效率。同时，有助于增强学生的学习兴趣和动机，进而提升学习效果。

2. 教育测评的方式从离散性到连续性。传统人工测评提供的是学生在特定时间点的某种特征或能力快照。虽然这些快照可以表明学生在特定时间点所知道和不知道的内容，但这些快照是孤立和零散的，通常只反映了学生当前表现，很难反应学生从过去到现在的变化。[①] 而人工智能教育测评可以嵌入到学生的学习过程之中，以非侵入性方式捕捉和模拟学生潜在学习、思维过程痕迹与图景，从而提供连续性的学生表现或成绩视图。可见，人工智能教育测评不仅提供学生的学习结果，还能揭示学生的学习过程。据此，师生能够有效地调整自己的教学或学习策略。

3. 教育测评的性质从统一性到个性化。传统人工测评项目或试题通常都是统一的，即无论学生的先前知识、能力、经验和文化背景如何，都会被分配到统一的任务或题目。有研究认为，这种将一套测评工具视为适合所有人的做法可能是不公平的，因为并非所有学生都有同样的知识和能力基础。而人工智能教育测评可以根据学生个体的能力、表现、兴趣、文化背景量身定制测评任务和项目，并生成个性化反馈意见

---

①Zachari Swiecki，Hassan Khosrav，etc. *Assessment in the age of artificial intelligence*，May 24，2022，https://www.researchgate.net/publication/360471346.

和测评报告。①

4.教育测评的指标从单维度到多维度。传统人工测评指标相对比较单一，往往只是考察学生掌握了哪些知识以及掌握这些知识的程度。而人工智能教育测评可以设计更多维的测评指标，利用增强的图形和模拟技术使测评项目更具响应性和情境化，如用户交互的日志数据或构建响应性自然语言，评估更复杂的学习成果，如解决问题技能、非认知能力等。

5.教育测评情境从教室到真实。在需要问题解决的测评中，传统人工测评情境往往不太真实。以医学专业学生为例，其测评任务主要是考察医学生是否具备或在多大程度上具备了诊断和治疗病人专业知识、专业能力和专业道德。但是，把医学生尤其是低年级医学生放在真实医院和真实病人情境中，可能过于危险。只能在教室笔试或面试，而教室的测评情境可能影响测评的有效性和可靠性。因为当真实的活动被转移到教室里时，其所处环境不可避免地发生了变化。它们变成了教室里的任务，成为学校文化的一部分，随后的学习、测评体系被完全封闭在自我证实的学校文化中。可是，在学校文化中学生取得的成功往往与真实情境表现不一致。② 而人工智能可以通过模拟或创建真实情境，检测学生在真实情境的表现。

除了以上优势，人工智能教育测评对所有学生提交信息使用统一的评分标准，有助于减少教师评分时的主观偏见和评分不一致。有研究者认为，人工智能教育测评可能是人工智能教育场景应用最具革新性的领域，前景广阔、潜力巨大。③ 但与此同时，也伴随着诸多的风险。

二、人工智能教育测评的风险

萨洛蒙等人将技术对学生的影响分为两种主要类型："技术本身"的影响和"使用技术"的影响。④ 由此，我们也可以将人工智能教育测评的风险分为"人工智能教育测评本身"的风险、"使用人工智能教育测评"的风险以及二者共同面临的风险。

（一）人工智能教育测评本身的风险

从"技术本身"来说，人工智能教育测评就是一种技术人工物。简单地说，所谓技术人工物就是既在技术中产生，也在技术中使用的物体，是人类使用的技术对象经过现象学还原后的物质（虚拟）客体。⑤ 荷兰技术哲学家皮特·克罗斯和梅莱斯认为，任何技术人工物都具有双重属性——物理属性和功能属性。一方面，物理属性是具有特定物质结构的物质体，它必须服从自然法则的统治。另一方面，技术人工物又具有功能这一本质属性。这意味着在人类行动情境中，该人工物能够被当作是实现某种目的的手段。易言之，该人工物是某种功能的载体。⑥ 据此，人工智能教育测评作为一种技术人工物，具有结构和功能双重属性。人工智能教育测评结构上的核心要素是数据和算法，服从算法法则。至于功能，人们之所以开发和使用人工智能教育测评，是因为它具有教育测评的功能——教育性或管理性功能，能满足某种教育和管理需求。对于教育测评来说，有效性和公平性是最重要的指标。那么在这些指标上，人工智能教育测评所面临的风险有哪些？

1.测评有效性和可靠性偏差。尽管有研究表明，一些自动化评分系统与人工评分高度相关（$r \approx 0.92$），而且自动化评分和人工评分在评估项目难度方面也呈现出高度相似性，这也是自动化评分有效性的

---

①Zachari Swiecki, Hassan Khosrav, etc. Assessment in the age of artificial intelligence, May 24, 2022, https://www.researchgate.net/publication/360471346.

②Zachari Swiecki, Hassan Khosrav, etc. *Assessment in the age of artificial intelligence*, May 24, 2022, https://www.researchgate.net/publication/360471346.

③José Carlos Sánchez-Prieto, etc. *AI-Driven Assessment of Students: Current Uses and Research Trends*, July 10, 2020, https://link.springer.com/chapter/10.1007/978-3-030-50513-4_22.

④Salomon, G., Perkins, D. N., & Globerson T. *Partners in cognition: Extending human intelligence with intelligent technologies*. Educational researcher, 1991, 20(3): 2 - 9.

⑤陈多闻：《技术使用的哲学探究》，沈阳：东北大学博士学位论文，2010年，第57页。

⑥陈多闻：《技术使用的哲学探究》，沈阳：东北大学博士学位论文，2010年，第60页。

证据。[①] 但不同的观点一直存在，有研究发现，自动化评分系统对不同族裔、文化背景和性别的学生论文评分有效性和可靠性存在显著差异。美国一款自动评分系统对非裔土味英语作文的评分明显低于对标准美式英语作文的评分。美国教育测试服务中心（ETS）的数据显示，女性研究生申请考试（GRE）平均得分比男性低 80 分。[②]

此外，在各类招生任务中，使用或辅助使用人工智能测评正变得越来越普遍。虽然这种方式简化了决策过程，减少了测评专家可能的偏见，但有时其有效性和可靠性也会引发人们的争议。

2020 年的春夏，英国政府把取消 A-level 常规考试作为对付新冠疫情的一种策略。但英国资格认证和考试管理办公室认为仅仅依靠教师的估计成绩不仅会导致分数膨胀，而且主观性也太强，有失公允。于是设计了一种自动化评分系统来评定学生的 A-level 分数，然而，2020 年 8 月，A-level 成绩公布后，引起了一场轩然大波。尽管相关部门声称："人工智能评定的总体成绩与前几年学生亲自参加考试的成绩大致相当，没有证据表明 2020 年人工智能评分过程引入了任何系统性偏见。"但 A-level 考生并不买账，他们认为人工智能系统对他们的评分与他们的平时成绩和教师估计成绩有不可解释的偏差。因为相比后两种分数，成千上万学生的分数被下调：35.6％的学生的分数被下调了一个等级，3.3％的学生下降了两个等级，0.2％的学生下降了三个等级。这意味着总共有近 40％学生成绩被降级。[③] 这个成绩直接影响到他们能否被心仪大学录取、能否获得奖学金以及进行学分互换。

无独有偶，由于同样的原因，2020 年国际文凭项目组织也使用人工智能来测评学生的成绩。2020 年 7 月成绩公布后，也引发了大规模的抗议和不满。大约两万四千学生，约占 2020 年修习国际文凭项目学生总数的 15％，签署了抗议书。国际文凭组织社交媒体页面上充满愤怒的评论，质疑智能测评结果的有效性和可靠性。[④]

2. 数据偏颇和算法偏见所导致的测评结果不公平。上述人工智能测评系统有效性和可靠性偏差可能源于系统收集和训练的数据偏颇及算法偏见，比如自动化英语评分系统的数据依赖的群体往往是说标准英语、中产阶层的男性。这样导致智能评分系统可能低估了少数族裔、女性和处境不利群体的语言结构，将语言习惯、文化或性别差异误评为语言能力的高低。易言之，人工智能教育测评的关键要素是数据和算法，如果收集和训练算法的数据不具有足够包容性和代表性，算法可能因数据偏颇或开发者无意识的偏见，导致测评结果出现系统性、可重复性偏差——可能特别有利于或者特别不利于某些群体，从而延续和加剧现有教育不公平现象。上述非裔和女性得分较低就是典型案例。2020 英国 A-level 成绩特别有利于学生人数较少私立学校的普通学生，部分原因是因为这些学校历年 A-level 成绩优秀；历史上成绩不佳的公立学校优等生得分则受到了限制。[⑤] 在国际文凭项目考试中，也有类似情况，美国科罗拉多州西敏斯特高中的伊莎贝尔的母语是西班牙语，在以往的西班牙语课上，连续两年获得 A＋的成绩，但这次人工智能系统给她测评的西班牙语成绩不及格，部分原因是她所在学校的历史上西班牙语成绩不佳。[⑥] 德克萨斯大学

---

①Jiangang Hao，Alina A. von Davier，etc. *Transforming Assessment：The Impacts and Implications of Large Language Models and Generative AI*，July 10，2020，https://onlinelibrary.wiley.com/doi/full/10.1111/emip.12602.

②John Gardner，Michael O'Leary，etc. *Artificial intelligence in educational assessment："Breakthrough? Or buncombe and ballyhoo?"*，July 5，2020 https://onlinelibrary.wiley.com/doi/full/10.1111/jcal.12577.

③Jon Porter. *UK ditches exam results generated by biased algorithm after student protests*，August 17，2020，https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications.

④Theodoros Evgeniou，David R. Hardoon，and Anton Ovchinnikov. *What Happens When AI is Used to Set Grades*? August 13，2020，https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades.

⑤Richardson，M.，Clesham，R. *Rise of the machines? The evolving role of AI technologies in high-stakes assessmen*，London Review of Education，2021，19（1）：1 - 13.

⑥Meredith Broussard. *When Algorithms Give Real Students Imaginary Grades*，September 8，2020，https://gcdi.commons.gc.cuny.edu/2020/10/14/opinion-when-algorithms-give-real-students-imaginary-grades-the-new-york-times/.

奥斯汀分校计算机科学系使用的“评分”项目也由于其算法过于依赖历史录取数据而备受指责。①

哪怕是在线监考系统，也可能由于考生肤色较深而无法被系统正确识别。阿姆斯特丹自由大学的诉讼案中，在线监考未能正确判断学生是否作弊，就是因为考生肤色问题。② 以上种种涉嫌数据歧视和算法偏见，延续甚至放大了社会对女性、有色人种和社会处境不利学生的歧视，限制了这些学生的平等权利和机会。

（二）使用人工智能教育测评相关的风险

1. 使用人工智能教育测评可能会限制测评的某些教学功能。人工智能教育测评不仅是一种测评工具，即衡量学生掌握了哪些知识、技能以及掌握程度。而且还是一种隐性课程，具备教学功能。比如，教师通常会利用传统人工评分形式来激励和引导学生，如果一个教师觉得某生会从鼓励中受益可能对该生的评分比较宽容。或者，教师觉得某生可能需要从干预中受益，评分就会更严谨。③ 这两种情况下评分都植根于教师对学生的了解以及教师与学生建立的个人关系。这两位学生不仅得到分数的反馈，还可能因体会到教师良苦用心，从而更加努力或更加严格要求自己。这就是人工测评的教学功能。而人工智能教育测评，可能会削弱教师对学生的全面了解，使测评过程过于机械和缺乏人性化，从而限制了某些教学功能的发挥。

2. 使用人工智能教育测评可能会带来某些教育性风险。使用人工智能测评还可能会出现出乎教育者预料、对学生隐性、负向的影响，我们可以将其称为“教育性风险”。如前所述的课业预警系统，就算技术上测评结果再准确，那些被预测到课业成绩可能不佳的学生，一个可能后果就是该生还没有开始学习课程就放弃了自己的努力，因为反正智能测评工具已经预测到自己课业成绩不佳。教师也可能因为看到名单而降低了对这些学生的期望，下意识地不愿意付出额外的努力。这是“罗森塔尔效应”的反向应用。还有，人工智能教育测评是非常精准化的，这种“精准化”可能会一定程度上限制了学生信息选择，产生“信息茧房效应”，影响学生的全面性发展。这都是使用人工智能教育测评或使用不当带来的教育性风险。而这种教育性风险通常是隐性的，因其内隐性和滞后性不易被发现。不仅如此，人工智能教育测评本身也是一种价值导向，暗含了一定的价值基础——关于教育是什么以及教育为什么的价值观。这些都是开发和使用人工智能教育测评所应该考虑的基本问题。

（三）人工智能教育测评本身和使用人工智能教育测评共同面临的风险

1. 可能会对学生造成侵权。首先，可能会侵犯学生的隐私权和个人数据保护权。隐私权和个人数据保护权在教育技术中并非新鲜话题。然而，人工智能教育测评由于需要海量数据，使得隐私权和个人数据保护权问题更加突显。人工智能教育测评数据通常是记录学生的课堂表现、成绩、所学知识和学习能力，但也可能包括他们的性格特征、偏好、健康状况、家庭情况以及其他有助于评估其能力的个人信息。尽管现有立法和标准旨在保护敏感的个人数据、保护个人隐私权。但人工智能技术公司由于利益驱使或技术原因，不断爆出数据访问的安全性事件，加剧了人们的担忧。④ 比如在线监考系统常常涉嫌侵犯学生隐私权和个人数据保护权。2020 年，荷兰乌特勒支大学的数据保护官因该大学使用在线监考软件的隐私问题与执行董事会产生了冲突。2021 年，华盛顿大学等一些知名高校宣布取消与一家知名在线监考公司续签合同。2022 年，美国俄亥俄州克利夫兰州立大学一名大学生赢得了联邦法官的支持，法官裁定“在线监考进行房间扫描违反宪法”。⑤

---

①Lilah Burke. *The Death and Life of an Admissions Algorithm*, December 13, 2020https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd.

②Sanderijn Kuijvenhoven. *The Quest for Ethical AI in Higher Education*, February 15, 2024https://studenttheses.uu.nl/handle/20.500.12932/46269.

③Zachari Swiecki, Hassan Khosrav, etc. *Assessment in the age of artificial intelligence*, May 24, 2022, https://www.researchgate.net/publication/360471346.

④N Smuha. *Trust worthy artificial intelligence in education: Pitfalls and pathways*, September 23, 2020, https://scholar.lanfanshu.cn/scholar?hl=en&q=TRUSTWORTHY+ARTIFICIAL+INTELLIGENCE+IN+EDUCATION%3A+PITFALLS+AND+PATHWAYS+

⑤Sanderijn Kuijvenhoven. *The Quest for Ethical AI in Higher Education*, February 15, 2024, https://studenttheses.uu.nl/handle/20.500.12932/46269.

其次，可能会侵犯学生的被公平评价权。如前所述，人工智能教育测评系统可能将种族、性别和社会文化背景编码到程序中，低估了女性、有色人种或者处境不利的学生，涉嫌性别、种族和社会背景歧视。英国 A-level 和国际文凭项目的考试中，有两个关键数据点是自动评分系统的依据，一个是考生所在学校历史成绩。一般来说，处境不利考生所在学校历史成绩往往表现不佳，哪怕考生本人很优秀。另一个数据点是教师估计的学生成绩，与白人学生相比，教师对黑人和棕色人种学生的期望往往较低。① 人工智能教育测评系统会放大这种偏见。

最后，人工智能教育测评系统可能会侵犯学生的知情同意权、版权或其他权利。比如，有学生在自动评分系统提交的论文在其毫不知情的情况下竟然成了某个平台的范文。这不仅侵犯了学生的知情同意权，还侵犯了学生的版权。

2.学生的主体性和自主性可能会被削弱。如前所述，人工智能测评系统的即时评分和反馈虽然有助于学生学习，但也可能让学生对系统产生依赖，削弱了学生自身的主体性和自主性。并且智能测评输出被自动视为正确更暗示了他们不需要独立思考和自主决策。不仅是学生，教师主体性和自主性都会受到影响。过去学生提交的作品是否与其他作品相似或雷同由教师来决定；而现在，由于可获得的资源数量庞大以及自然语言处理技术进步，教师更依赖于人工智能处理这项任务。② 这样，教师对智能测评系统的依赖又以隐性课程方式微妙地影响了学生的自主性。

三、人工智能教育测评风险规避框架

综上可见，开发和使用人工智能教育测评有明显优势，但也伴随着诸多的风险，那么怎样规避这些风险呢？图 1 提供了人工智能教育测评的风险规避框架。这个框架要求开发和使用人工智能教育测评首先要评估这一系统对学生的益处及可能的风险或可能对学生的伤害。其次，在测评的任何环节，都应该保障学生各种权利，避免侵犯学生权利。当然，这些要求实现的前提是人工智能教育测评开发和使用必须以人为中心，以尊重和补充人尤其是学生的主体性和自主性为宗旨。最后，这些要求实现的保障是可问责性，一旦出现对学生的侵权和风险，相关人员能够被追责。以下是具体的分析：

（一）人应该是人工智能教育测评的监督者和最终决策者

人工智能必须以人类为中心，开发和使用人工智能测评系统的任何环节都需要人类监督，最终决策应该由人来做出。2020 年的国际文凭项目和英国 A-level 考试人工智能测评风波之所以引起这么大的愤怒和抗议，第一，面对考生质疑和申诉，不管是国际文凭组织还是英国当局起初都坚持认为人工智能测评是准确的，过于信赖和依靠机器。第二，在国际文凭项目和英国的 A-level 人工智能评分过程中和公布成绩之后，相关人员没有发挥有效的监督和调整作用。如国际文凭项目成绩申诉程序并没有因为测评系统变成人工智能而做出针对性调整。面对申诉国际文凭组织没有预案，他们本应该对申诉人的异常分数进行由人主导的重新评估，审慎评估他们输入的数据，并研究如何系统解决问题。③ 但显然国际文凭项目组织并没有这么做。英国的 A-level 成绩申诉程序也没有因评分方式转变进行有效性调整。考生们被告知他们可以用模拟考试成绩作为申诉的依据。然而，因为疫情有些学校并没有组织模拟考试，没有模拟考试成绩的学生怎么申诉？申诉主体是学生还是学校？费用由谁承担？④ 这些规定都非常不明确，加剧了考生及家长对人工智能评定成绩的不满。第三，对于学生利害攸关的考试成绩由自动评分系统评定，涉嫌违法。不少人认

---

①Meredith Broussard. *When Algorithms Give Real Students Imaginary Grades*, October 14, 2020, https://gcdi.commons.gc.cuny.edu/2020/10/14/opinion-when-algorithms-give-real-students-imaginary-grades-the-new-york-times/.

②Zachari Swiecki, Hassan Khosrav, etc. *Assessment in the age of artificial intelligence*, May 24, 2022, https://www.researchgate.net/publication/360471346.

③Theodoros Evgeniou, David R. Hardoon, and Anton Ovchinnikov. *What Happens When AI is Used to Set Grades?* August 13, 2020, https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades.

④Avi Gesser. *The UK School Algorithm Debacle: Five Lessons for Corporate AI Programs*, August 25, 2020, https://www.debevoise.com/insights/publications/2020/08/the-uk-school-algorithm-debacle-five-lessons-or.

为，2020年国际文凭项目人工智能评分有违欧盟的《通用数据保护条例》有关条款。“数据主体有权不受仅仅基于自动处理（包括分析）而做出的决定约束，如果该决定对数据主体产生法律效力或类似的重大影响。”① 其实，在A-level成绩公布的几个月前，英国皇家统计学会曾对英国资格认证和考试管理办公室发出警告：“永远不可能有一个完美的算法或一个完全公平的系统，必须有健全的人类决策，否则可能是一场惨败”。② 但英国当局对此没有做出回应。可见，在开发、使用人工智能测评的每个环节都需要建立由人主导的评估机制，以监测人工智能测评的过程、结果以及对学生的影响，并根据监测进行相应的调整。

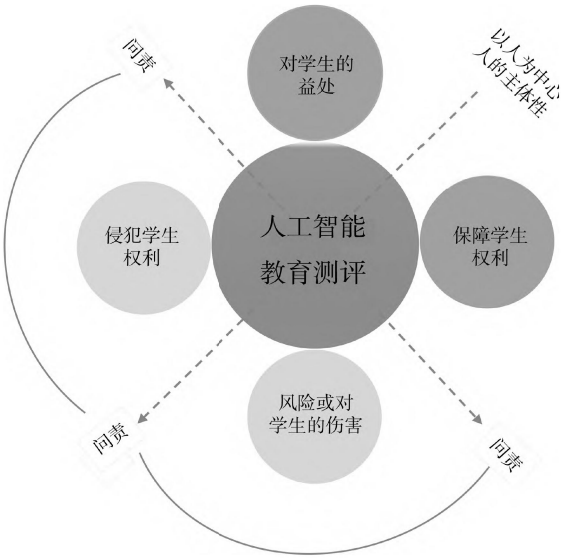


图1　人工智能教育测评风险规避框架

（二）人工智能教育测评开发、使用的任何层面，学生主体性和自主性必须被尊重

随着人工智能自主控制和自主决策能力不断增强，越来越具有拟主体的身份。为了避免人类主体性被人工智能反噬带来不可控制的风险，在开发和使用人工智能过程中，人作为人的自主性、主体性必须被尊重。2019年4月欧盟颁布《可信赖的人工智能伦理准则》，把“尊重人类自主性”视为人工智能开发和使用的首要伦理准则。2019年5月，北京智源人工智能研究院联合北京大学、清华大学、中国科学院自动化研究所、中国科学院计算技术研究所、新一代人工智能产业技术创新战略联盟等高校、科研院所和产业联盟共同发布《人工智能北京共识》，在“服务于人”条款中明确提出“应充分尊重人类的隐私、尊严、自由、自主、权利”。其他国家或者相关人工智能组织也都发布过类似伦理准则。具体到人工智能教育测评系统，其预设前提、价值基础都需要把人的主体性尤其是学生的主体性放在优先位置。以自动化作业批改系统为例，其价值预设应该是培养学生的自我反思能力和元认知能力，鼓励学习者本人反思、研究自己的答案究竟错在哪里？为什么会出现错误？而不是养成学生完全依赖自动化作业批改系统来修正答案的习惯。而对于教育者来说，人工智能测评教育功能的发挥不仅依赖智能系统本身的结构数据和算法，而且也依赖于师生主体性的努力。如前所述的人工智能预测学生成绩及其辍学率，既可能造成师生不愿意努力而放弃学生学业的“罗森塔尔效应”反向应用；也可以通师生的主动性努力将这种负面影响转化为正面影响。那些被预测到学业可能受挫的学生，他们提前知晓了学习难度，学生选择更加刻苦学习，教师也针对这些学生提供额外支持和鼓励。这样，通过师生的主观性努力，人工智能预测所可能带来的消极影响就变成了对学生积极影响。还有，人工智能教育测评的精准化、个性化带来的“信息茧房效应”，也可以通过教师有意识地辅导和

①*EU data protection rules*，December 24，2020，https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en.

②Cath Everett. *UK A-level algorithm fiasco a global example of what not to do-what went wrong and why*，August 24，2020，https://diginomica.com/uk-level-algorithm-fiasco-global-example-do-wrong-and-why.

学生补充练习在很大程度上预防和消解。

（三）开发和使用人工智能教育测评需要遵循审慎原则，评估其益处和风险

从教育学的角度来看，教育测评直接功能主要有两个：第一，教育教学功能，即为了促进学生的学习和全面发展。第二，行政管理功能，通常表现为利用测评区别学生水平，给予适当安置或选拔淘汰；或为提高行政效率或减轻教师负担。当然教育测评的终极追求依然是教育性，管理功能最终需要符合教育功能。人们开发和使用人工智能教育测评系统初衷是期待发挥上述两个或一个功能的发挥。如前所述的德克萨斯大学奥斯汀分校计算机科学系采取“评分”项目主要是为了节省该系招生委员会成员的时间，减轻招生教师负担，从而使这些教师有更多的精力和时间服务于学生。

不管哪一种情况，开发和使用人工智能教育测评首先需要考量它作为测评工具的有效性、可靠性和公平性等指标，评估它对学生的益处。与此同时，还需要评估可能对学生造成的负面影响，比如一款智能测评软件可以提高学生的学习兴趣和学习效率而对学生有益。但学生长期沉湎于人机互动，可能造成真实世界的情感冷漠，那么在开发和使用过程中需要辅以哪些措施才能避免这种情况发生等诸如此类问题，都需要充分讨论和监测。哪怕初衷是行政管理或者不得已而为之的情形，也需要考虑会不会对学生造成负面影响，需要采取哪些措施去规避。易言之，开发和使用人工智能教育测评需要遵循审慎原则，评估收益的同时进行充分的风险评估、风险管理和风险控制。其实，2020 年英国的 A-Level 考试，英国皇家统计学会曾向英国资格证和考试管理办公室发出过警告，说用于测评成绩的算法不稳定，风险很大。① 但该机构置若罔闻，还是继续推进自动化评分。先不说这些利害攸关的教育测评对学生的影响是关键性的，影响其升学、就业甚至改变一生的命运。即使平时测评对学生的影响也是不可低估的，所以最大限度地发挥人工智能教育测评对学生的益处、最小化对学生的负面影响至关重要。

（四）人工智能教育测评开发和使用要保证学生的权利不被侵犯

如前所述，人工智能教育测评需要海量的数据支持，而很多数据与个人信息和隐私权相关联。而且怎样使用数据、使用数据的性质（依赖于算法）都可能会侵犯学生隐私权、被公正评价权、版权等其他权利。虽然，面对飞速发展的人工智能技术和铺天盖地的数据，各个国家和地区纷纷制定法律和标准以保护个人基本权利。但是，人工智能教育测评侵犯学生权利的事件屡见不鲜。为了解决这些问题，在训练人工智能教育测评系统之前，开发人员必须采取措施保证数据的包容性和代表性，避免数据歧视和算法偏见。确保数据的机密性、完整性和可用性以及防止数据泄露。同时，部署和使用人工智能教育测评的机构和个人负有同样重要的伦理责任。因为数据不仅在智能测评系统开发阶段至关重要，在使用过程中也同样重要，系统运行使用过程中产生的数据甚至需要作为新的输入再次使用。所以在开发和使用人工智能教育测评的任何环节都需要明确规定和严格执行访问权限，防止未经授权的访问。此外，学生知情同意权和的版权等其他权利都需要被保护，明确人工智能教育测评数据信息在整个生命周期都不能用于规定之外的功能，最大限度避免对学生权利的侵犯。

（五）人工智能教育测评系统的算法要稳健、透明

算法是人工智能教育测评的核心结构之一，它指导计算机如何处理数据、学习模式、做出决策和生成输出。为了规避相应风险，对人工智能教育测评算法的要求通常有两个方面，一方面是稳健，要求算法准确、可靠、可重复，避免算法的错误、算法不一致或不稳定带来的风险。显然，英国 2020 年 A-level 人工智能评分系统的算法是不一致的。虽然大部分学生的成绩依据是学校的历史表现和教师的估计成绩，但如果一所学校选考某一学科的学生少于五个，那么他们的成绩只根据老师的估计成绩来决定。如果某一学校选考某门学科的学生有五到十个，教师的估计成绩仍然会被赋予更多的权重。私立学校人数相对较少，这也是导致私立学校学生得分较高的一个原因。② 另一个方面算法透明，表现为可解释、可追溯。可解释意味

---

①Cath Everett. *UK A-level algorithm fiasco a global example of what not to do-what went wrong and why*, August 24, 2020, https://diginomica.com/uk-level-algorithm-fiasco-global-example-do-wrong-and-why.

②Sky news. *A-levels and GCSEs: How did the exam algorithm work?* August 20, 2020, https://uk publiccollege.co.uk/blog/a-levels-and-gcses-how-did-the-exam-algorithm-work/.

着人工智能教育测评系统的工作原理、工作机制都是可以被解释的。可追溯指的是人工智能教育测评系统做出决策的所有基础，包括收集数据、标记数据、使用算法等过程都应该被详细、系统记录，即系统所做出的任何决策都是有迹可循的。[①] 这样一旦出现问题，从算法上可以追溯到原因以实现问责。

（六）保证人工智能教育测评的可问责性

可问责性既是对上述要求的重申，也是上述要求的保证。同时，将教育测评权委托给人工智能本身就是一种人类选择，属于人类的责任范畴。这意味者开发、使用人工智能教育测评所导致的结果或后果需要由人类承担，而不能推给机器，尤其是是在对学生造成不利影响时相关人员必须承担责任。为了保证测评系统的可问责性，首先，部署和使用人工智能测评前权衡，评估测评的收益和风险，并明确权衡主体的责任。其次，使用过程中或使用后的评估，人工智能测评系统的整个生命周期需要由相关人员审核评估，评估师生或其他管理人员的使用情况以及后果，尤其是可能面临风险或者可能对学生伤害时的责任承担。比如某款人工智能测评系统有效性、公平性有偏差或者对学生产生了不利影响，需要确定到底是系统开发者算法设计上的失误、偏见？数据收集和训练中的歧视或包容性不够？或者是学校在部署选用系统时没有进行充分的评估？还是教师在使用过程中没有进行适当的引导、监督或干预？或者是属于学生操作不当？责任归属如何划分，责任范围如何界定，相关责任如何追究都需要明确的规定和监督。最后，可问责性还包括在人工智能教育测评的任何环节，一旦涉嫌对学生造成了伤害或者侵权，都应该有迅速及时的补救措施。

人工智能教育测评的风险规避不会一蹴而就地解决，因为这些问题不仅事关教育测评的理念、相关人员的认知，还涉及人工智能技术难度以及各个要素之间的张力。但这并不能否认解决这些问题的紧迫性，本文希望对这些问题的讨论能唤起人们对人工智能教育测评应用、风险及规避充分的关注和重视，这事关人工智能教育测评的未来，也事关教育的未来。

International Perspective on Artificial Intelligence Education Assessment: Application, Risks, and Mitigation

XU Xiang-he

(School of Education Science, Liaocheng University, Liaocheng 252000, China)

Abstract: Artificial intelligence assessment in education is a trend in the digital age for the development of educational assessment, including automated scoring systems, computer adaptive testing, process-based formative assessment and prediction, and integrity assurance mechanisms. Compared with traditional manual assessment, the advantages of artificial intelligence assessment in education are very obvious. However, it also faces many risks, specifically the risks of artificial intelligence assessment itself, the use of artificial intelligence assessment, and the risks common to both. In order to mitigate these risks, firstly, the design and use of artificial intelligence assessment in education need to assess its risks and benefits to students; secondly, privacy and other rights of students should be protected at all stages of assessment. Of course, these requirements can be achieved only if artificial intelligence assessment in education is centered on humans and aims to attach importance to the agency and autonomy of students. If there is any infringement or harm to students, relevant personnel can be held accountable and there should be timely remedial measures.

Key words: artificial intelligence assessment in education; risks; mitigation; international perspective

[责任编辑　治平]

①Ryan S. Baker, Aaron Hawn. *Algorithmic Bias in Education*, International Journal of Artificial Intelligence in Education, 2022, 32(3): 1052 - 1092.