论算法应用于未成年犯人身危险性评估研究

赵桐玥

上海师范大学

摘　要：近年来，随着人工智能的不断发展和大数据的广泛运用，算法技术逐渐渗透至刑事司法领域。由于未成年犯的特殊性，对其进行人身危险性评估极为必要，但算法的应用也存在算法歧视、算法黑箱、专业术语的理解不足、缺乏经验性的问题。构建未成年犯人身危险性评估算法系统应当保障数据的准确性，赋予评估对象算法解释权，建立专业语言数据库，强调评估结果的可追溯性及其参考性。

关键词：算法；未成年犯；人身危险性评估

【DOI】10.12252/j.issn.2096-627X.2023.06.221

一、算法在司法领域的应用现状

（一）算法在国内司法实践中的应用

近年来，随着人工智能的不断发展，以大数据为载体的算法时代拉开序幕。在国家政策的指导下，算法在司法领域中的应用范围不断扩大。

北京市高级人民法院研发推出了“睿法官”系统。该系统通过对不同案件进行学习，增强自身知识的全面性和系统性，能够自动提取与案件有关的信息，根据当事人的情况及案件的信息进行自动匹配，进而推荐情节相似的案例，[①]根据已经得到的信息智能生成裁判文书。

上海市高级人民法院推出了“206”刑事案件智能辅助办案系统。这一系统能够进行语音识别、自动抓取与示证，具有类案推送、文书自动生成、办案人员知识索引等功能，能在繁杂的卷宗中提取有用的信息，帮助法官提升办案质量和效率。2017年“206”系统投入运行，将刑法中的故意杀人罪、盗窃罪、非法吸收公众存款罪和诈骗罪纳入其中。

（二）算法在国外司法实践中的应用

美国已经推出了以替代性制裁为目标的惩教犯管理画像（COMPAS）、公共安全评估（PSA）和服务级别目录修订版（LSI-R）三大风险评估软件，主要用于刑事诉讼中的司法审判环节，可以预测罪犯再次犯罪的风险性、开庭日出庭的可能性等，并在结果上影响关于量刑、保释和假释的决定。2016年，英国研发了一种新的人工智能系统，能够帮助司法人员处理司法案件、预测案件裁判结果。

二、算法应用于未成年犯人身危险性评估的可行性分析

（一）评估对象可接受性增强

英美等国家已经研发出以算法大数据为基础，预测个别罪犯是否有再犯可能性的风险评估软件，被广泛应用于审判前风险评估、量刑、缓刑和假释工作中，主要包括COMPAS、PSA、LSI-R等。[①]评估对象对于软件的接受性能够高于面对司法工作人员进行测评的接受性。未成年犯在面对成年人时，潜意识中会有一种隐藏的自我保护意识。将算法应用于人身危险性评估中，受评估对象不用与司法工作人员进行交流，只需要按照软件的提示进行作答，就能够完成人身危险性的评估。避免了未成年人拒绝交流这一情况的出现。利用算法软件进行评估更能提高未成年犯的配合性。评估全过程是由算法自主完成的，工作人员只需对评估结果进行审查即可。

（二）评估过程更加客观

算法系统拥有机器学习功能，能够让算法通过学习不断提高自身的性能，机器学习功能也从浅层学习向深层学习发展。人身危险性评估软件可以筛选出影响人身危险性程度的相关数据，分析得出未成年犯再犯可能性，得出结论。这一过程并没有司法工作人员的参与，未成年犯直接通过算法软件进行人身危险性的评估。摒弃掉评估人员自身对于各类犯罪、性别、地域或自身经历带来的一些偏见，减少这种主观倾向带来的误差，最大程度上保障信息的真实性，得到更加客观的评估结果。通过算法系统进行评估，受评估的未成年犯适用统一的标准，评估结果能够准确反映犯罪人的人身危险性，对于受评估的未成年犯来说这样的结果更加公平合理。

算法能够吸收法学、心理学、社会学等相关的知识理论，更加全面地对未成年犯进行人身危险性评估，综合分析各项影响因素，得出专业、客观的评估结论。弥补司法工作人员、社区民警等这些传统评估者的不足，提高人身危险性评估的专业性。

（三）评估具有更强的适应性

算法相较于人类搜集信息来说，能够打破人身危险性评估时效性和地域性的限制。未成年犯心智并不成熟，认知能力在不断提高，其人身危险性也会随着其对社会的认识、思考、理解而不断变化。评估涉及的相关因素并不是静态的，而是动态的。这些据以作为依据的相关因素只有根据社会、政策、舆情的变化而变化，才能够得出具有实用性的评估结果。传统的测评表通过抽样的方式对相关危险性因素进行选取和评价，是依据对特定时期、特定地域或特定罪名的未成年犯的分析而设计，当情况发生变化时，传统测评表并不能做出相应的改变，因此难以保证其可信度、有效性及准确性。[②]但算法的机器学习功能能够及时对相关因素进行更新，优化模型，生成新的算法，得出最新的、有效的结论。

不同地区的经济发展水平不同，在一定程度上也会

影响该地区的犯罪率、犯罪类型、公众的法律意识及未成年人的受教育程度。经济发达的地区相较于不发达的地区，犯罪率相对较低，经济类型犯罪相对较少，公众具有较强的法律意识，对于法律的规定有基本的认知。经济水平较高的地区教育资源丰富，未成年人能够更好地进行学习，对社会的认识程度更高，能够独立进行理性的思考。经济发展水平对相关危险性因素有着潜移默化的影响，所以人身危险性评估的标准应当注意到地域的区别。不同地区内传统的风俗习惯也会对未成年人产生影响，是相关危险性因素之一。算法系统通过自身学习，使得评估能够随着地域经济水平、传统习俗、社会变迁等情况实时更新。

（四）评估结果准确性提高

算法具有技术理性，能够精准把握未成年犯人身危险性评估中那些风险的成因、比例权重等影响因素，通过算法得出的评估结果更加准确。将算法应用于未成年犯人身危险性评估中，以大量数据作为基础，将所有信息转换为代码，依据设定的模型进行计算。全过程排除了人为因素的干扰，考虑的信息较为全面，评估结果相较于司法工作人员得出的结论更为精准。

在人身危险性评估过程中能够发挥算法的优势，将繁冗复杂的信息和数据利用算法进行规制整理，针对不同的未成年犯的不同的经历和心理状态进行有针对性的评估，解决不同主体间的差异性问题。根据未成年犯自身的独特经历，对影响人身危险性评估的相关因素赋予不同的权重，这样针对未成年犯具有独特性的评估结果更为准确且有说服力。

（五）基础数据全面

我国现有的未成年犯人身危险性评估有三种载体，分别为心理专家出具的心理测评报告、社会调查报告、以心理学为基础的各类测评量表。[①]这三种载体在具体案例中互为补充，但这三种方式获得的信息并没有有效整合的路径，降低了人身危险性评估的效率，影响了评估的准确性和全面性。

应用算法进行评估，获得数据的途径主要是通过大数据平台和专业数据库，数据广泛且全面，包含社会生活中各类信息，如电话、电商平台地址、消费记录、网站浏览记录、出行信息、社交账号等。同时数据的来源不仅包含中国裁判文书网这样的司法数据，还有来自其他行政部门或电商平台的生活数据信息。全面、丰富的信息在算法模型的运行下推测出事物的全貌的可能性大大提升，在很大程度上节省了人力，提高了工作效率。

三、算法应用于未成年犯人身危险性评估存在的问题

将算法应用于未成年犯的人身危险性评估中不仅获得了大量基础数据，而且使得评估过程客观、具有较强的适应性，评估结果更加精准，同时也提高了评估对象的可接受性。但是算法作为一种技术，具有其自身的局限性，将其引入到未成年犯人身危险性评估中，暴露出以下问题：

（一）算法歧视

歧视是人类认知世界的副产品。尽管是在人工智能时代，具有学习能力的机器也并不能排除存在歧视的情况。算法是由代码所组成，这些代码是由程序员编写的，隐藏在算法程序背后的歧视更加隐蔽，难以被人们所察觉，但这并不代表不存在歧视。美国人工智能风险评估软件COMPAS被指出屡屡在再犯风险评估中涉嫌种族歧视。算法歧视并不是技术手段创新所导致，而是数字时代前沿科技“赋能”传统社会矛盾的结果。

算法歧视并非技术创新的产物，而是数字时代前沿科技“赋能”前数字时代传统社会矛盾的结果。[①]算法虽然是客观的，但是算法的开发者是具有主观偏见性的，在编写代码的过程中会不自觉地加入自身的价值观，无意识的歧视开始滋长。大数据本身并不是完全保持中立的态度，其背后所反映的是人类社会存在的歧视和不平等。

（二）算法黑箱

算法并不是透明的。基于算法的专业性和技术性，不具有计算机相关知识的人无法理解其具体的运行过程，无法判断是否存在歧视、区别对待等问题。美国“威斯康星州诉卢米斯案”中，由于法院无法得知案件中对被告人所使用的风险评估软件COMPAS的具体评估方法，并没有判定该系统存在歧视。系统开发者利用商业秘密保护算法，使得内部的运转神秘、难以捉摸。算法所搜集的数据中有的与国家秘密、商业秘密及个人隐私有关，在对未成年犯进行人身危险性评估时，这些信息并不能公开，不能将其所涉及的风险比重等问题做出详细的解释。算法黑箱就此产生。算法黑箱限制了评估对象的知情和抗辩的权利，影响评估结果的说服力、可采性。

（三）不能充分理解专业术语

语言的语法和语义较为复杂，算法并不能充分理解各个专业领域内的术语。虽然可以通过代码将各类术语编写进算法程序中，但代码不能处理带有感情色彩的词语和句子。对于未成年犯的人身危险性评估涉及多门学科，但最主要还是法律的相关概念及规定。即使是法官，对于法律概念和法律规则的理解也是具有差异的，也要通过法律解释等法律方法对条文进行学习。对未成年犯进行人身危险性评估，不仅是通过技术分析得到简单的结果，更重要的是在充分理解各领域的前提下做出合法、合理的评估。

（四）缺乏经验性

在未成年犯人身危险性评估中，算法在理解性方面具有局限性，相对于工作人员，其运行过程缺乏处理未成年案件的经验。算法根据设定好的程序运行，依照法律、法规等规范性文件及相关风险因素，在各类数据中自动抓取信息，生成评估结果。算法机械运行，严格按照编写的代码排除各种人为因素的干扰。这样生成的结果虽然客观，但缺少了司法工作人员的经验和思想，难以实现法理与情理的统一。算法依靠设计好的算法和纯粹的理性逻辑进行分析，其输出结果往往是固定的标准答案。

四、算法应用于未成年犯人身危险性评估的设想

（一）保障数据的准确性

在采集数据的阶段，应当确保数据的准确性和全面性，从根本上遏制虚假数据、数据代表性过度、数据代表性不足等问题的出现，这样才能保证算法运行所生成的结果准确有效。尤其是在数据收集量不足时，数据的准确性和全面性尤为重要。未成年人本身就存在大量不确定因素，如年龄、性格、环境等，这些都影响对其进行再犯风险的分析。评估结果作为减刑、假释、保释的依据，只有保障源头数据的真实准确，评估结果才能具有说服力，维护社会的公平正义。

（二）赋予算法解释权

赋予未成年犯算法解释权，能够更好地保障未成年人的权益，使其能够更好地理解算法运行的过程，增强人身危险性评估的说服力。但算法解释也并不是披露算法技术的细节，只是将算法的运行逻辑、输入数据的含义、风险评估的考量因素及方法等对测评对象进行解释。这样既可以保护商业秘密，又可以增进测评对象对算法的理解，避免解释缺乏实际意义。赋予评估对象算法解释权可以帮助其了解算法，理解算法决策的逻辑，提高算法的透明度。

（三）建立专业语言数据库

随着人身危险性评估相关理论及其在司法实践中的应用的不断发展，涉及的专业知识领域愈加复杂，包含法学、心理学、犯罪学、社会学等多项学科。算法在人身危险性评估中缺乏专业语言。因此需要着力研发专业信息资源，全面搜集整合专业术语，利用技术优势，建立专业语言数据库。精进算法的语言结构，进而提高算法技术对于专业术语的理解、应用和分析水平，增强其在司法实践中的精准度。

此外，人身危险性评估所依据的数据多是用于刑事诉讼或刑罚矫正的司法数据，涉及定罪、刑罚的该内容，存在一定的歧视色彩。建立人身危险性评估的专业数据库，有助于在收集数据信息过程中，避免算法简单学习其中的偏见，而是能够在数据中客观提取有效信息，进行整理分析，最终输出结论。

（四）强调评估结果的可追溯性

将算法应用于未成年犯人身危险性评估的基本运作逻辑就是依据搜集录入的大数据，通过算法程序，得到最终的评估结果。但在这一过程中由于算法的专业性、技术性及算法的不透明性形成了算法黑箱，这与司法的公正公开要求相违背。有的司法工作人员为了规避自身的责任，可能会全盘接受算法所生成的评估结论。由于评估过程的不透明，这一过程所带来的结果就会受到各方的质疑，影响司法权威性及公信力。虽然算法是机械的机器运行，但人身危险性评估的数据、算法的设计等都需要人类参与，其中就会隐藏着人类的偏见。应重视评估结果的可追溯性，保证司法人员能够对人工智能的评估依据、评估过程进行分析和解释，对评估结果进行审查、修正。

（五）明确评估结果的参考性

目前，社会公众对于算法的信任远不及对于司法工作人员的信任。司法工作人员在评估过程中应处于主体地位，具有最终决定权。要明确算法在未成年犯人身危险性评估中的辅助地位，可以参考算法技术对于未成年犯的信息及评估结果。算法仅具有工具性价值，真正做出假释、减刑等决定的永远都是法官，并不会被任何技术所替代。

通过算法所得出的评估结果仅具有参考性，司法工作人员不能过分依赖算法进行的评估，应当根据具体的案情和未成年人的具体情况进行分析，做出最终的判断。未成年犯人身危险性评估并不是算法和人工审核的简单结合。将算法引入人身危险性评估中，目的是发挥算法技术的优势，处理繁冗复杂的琐碎性工作，减轻司法工作人员的负担，使其将重心放在算法不能操作的工作中。尊重现行的智能司法改革并不是一味地追求效率，而是利用算法技术的优势与创新不断优化原有的行为模式，在未成年犯危险性评估中实现公平正义。

参考文献

[1]申纯.人工智能时代人身危险性评估发展的新机遇及实现路径[J].求索，2021（06）：174-181.

[2]马国富，王子贤，马胜利.机器学习模型在预测服刑人员再犯罪危险性中的效用分析[J].河北大学学报（自然科学版），2017，37（04）：426-433.

[3]李棒.论人工智能技术在涉罪未成年人人身危险性评估中的应用[J].预防青少年犯罪研究，2018（06）：55-60.

[4]李成.人工智能歧视的法律治理[J].中国法学，2021（02）：127-147.

[5]陈光中.刑事诉讼法（第六版）[M].北京大学出版社，2016.

[6]于淼，陆娇，管政翔等.人工智能在司法量刑中的应用沿革与技术演进[J].西北工业大学学报（社会科学版），2021（03）：88-95.

[7]赵杨.人工智能时代的司法信任及其构建[J].华东政法大学学报，2021，24（04）：73-82.

[8]卞建林，曹璨.信息化时代刑事诉讼面临的挑战与应对[J].吉首大学学报（社会科学版），2021，42（05）：23-30.

[9]程龙.人工智能辅助量刑的问题与出路[J].西北大学学报（哲学社会科学版），2021，51（06）：163-174.

[10]狄小华.再犯风险评估的智能化研发研究[J].司法智库，2020，2（01）：14-27.

[11]卫晨曙.美国刑事司法人工智能应用介评[J].山西警察学院学报，2020，28（04）：22-28.

[12]朱晖，邵靖璇.AI在未成年人犯罪风险防范中的应用[J].南海法学，2020，4（06）：81-90.

作者简介：赵桐玥（1999—），女，汉族，天津人，上海师范大学在读研究生，研究方向：刑事诉讼法。