NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2025/04/04/unpriced-risk-is-not-return-why-most-artificial-intelligence-business-cases-in-human-resources-fail-as-business-cases/.

未定价风险不是回报:为何人力资源领域的大多数人工智能商业案例作为商业案例都是失败的

Heather Grizzle · April 4, 2025

The measurable savings in HR AI are small and the large ones cannot be measured. What a defensible AI business case in human resources includes.

目录

1. 需要哪些证据主张,以及支撑这些主张所需的测量方法
2. 真正可测量的是什么,以及它的价值几何
3. 录用质量并非一项指标
4. 预测性离职与准确率差距
5. 自动化合规并不能免除合规义务
6. 模型所遗漏的成本
7. 一种站得住脚的结构
8. 结语
9. 参考文献

需要哪些证据主张,以及支撑这些主张所需的测量方法

人力资源领域人工智能的标准商业论证已围绕五项主张趋于稳定:更快更好的招聘、更低的行政成本、可预测的人力规划、更高的留任率,以及自动化合规。这些主张在供应商材料、会议议程和专业评论中被反复重申,一致性之高已使其从需要证据支持的命题,变成了背景假设。

它们需要证据支持,而产生这些证据的障碍并非技术本身。人力资源部门进入这一阶段时,其测量基础设施在所有主要企业职能中最为薄弱。财务部门有经审计的报表。运营

部门有产出量和缺陷率。人力资源部门所使用的一系列指标定义不一,鲜少经过验证,在实施干预措施之前很少设定基线,而且往往由正在被评估其产品的供应商自行提供。一个在部署自动化工具之前都无法衡量其招聘流程绩效的组织,在部署之后同样无法衡量其改善程度,而它所报告的所谓回报,通常只是两个从未具备可比性的数字之间的差值。

Novara Consulting Group's position is that the central problem with artificial intelligence business cases in human resources is not that the returns are absent. Some are real, and they are described below. It is that returns are asserted at a level of precision the underlying measurement cannot support, while the costs that would offset them are excluded from the model entirely. A calculation constructed that way is not an investment analysis. It is an advocacy document with a currency symbol in it.

真正可测量的是什么,以及它的价值几何

这是一份印着货币符号的宣传文件。

五项标准主张可以清晰地分为两类,而这种划分本身比这些主张更具启发性。

第一类效应确实可测量,但相对而言效果有限。招聘周期是可测量的,大多数申请人跟踪系统中都已建立基线,并且确实会因排期、筛选和通信环节的自动化而有所改善。薪酬与福利管理中每笔交易的行政成本是可测量的,这些流程中的错误率同样可测。实现这些职能自动化的组织通常确实能够实现节省,而且这些节省是可以证实的。

这里有两点需要说明。第一,薪酬和福利管理中已记录的节省绝大多数来自工作流自动化和基于规则的处理,这些技术比当前这一代人工智能早了几十年,而在供应商的定位话术和据此构建的商业论证中,却常被重新贴上“人工智能”的标签。将早已存在的流程自动化所带来的节省归功于近期采购的人工智能,不仅夸大了此次采购所衡量出的回报,更重要的是,这错误地归因了因果机制,这意味着组织对究竟是什么产生了收益得出了错误的结论。第二点说明是,这些节省从绝对数值来看并不惊人。它们本身并不足以证明当前企业所做出的这些重大投入是合理的。

第二类效应如果属实将会相当可观,但目前无法测量:更优的人才录用质量、更高的留任率,以及更好的人力规划。正是这些主张撑起了整个商业论证,而它们之所以能够撑起论证,恰恰是因为其量级不受测量的约束。

录用质量并非一项指标

录用质量几乎是每一份招聘技术商业论证中最具支撑作用的主张,而它并没有一个经过验证的通用定义。根据不同组织,它被操作化为首年绩效评级、九十天留任率、十二个月留任率、招聘经理满意度、达到生产力所需时间,或以上若干项的复合指标,每一种都有不同的测量属性,而没有一种作为招聘有效性的衡量标准建立了公认的构念效度。

其结果是,任何所宣称的录用质量提升在很大程度上只是定义选择所产生的人为效应。一个将录用质量定义为九十天留任率的组织,若部署了一个筛选出与既往长期在职员工相似候选人的工具,就会观察到提升。它并未由此得知招聘变好了。它得知的是:一个针对代理指标优化的系统推动了该代理指标的变化。而在供应商同时提供工具和测量框架这一常见情形下,所得出的结论更接近同义反复,而非证据。

这一点的重要性远超出方法论上的严谨性。根据29 C.F.R. 第1607部分《雇员甄选程序统一指南》,若雇主的甄选流程显示出不利影响,该雇主须提供证据证明该程序对该职位而言是有效的。基于标准的效度验证需要一个站得住脚的工作绩效衡量标准。一个连自身职位的"录用质量"究竟意味着什么都无法精确说明、也无法证明其标准衡量指标本身不带偏见的组织,并不具备该框架所要求的证据。放大商业论证的测量薄弱之处,与削弱法律辩护的薄弱之处是同一个薄弱之处,这是一个方法论问题与法律责任问题恰好合而为一的罕见情形。

预测性离职与准确率差距

预测性人力规划值得单独讨论,因为已发表的性能数据与实际部署的性能之间的差距在这里最大,而其法律风险也最未被理解。

The academic literature on employee attrition prediction routinely reports accuracy figures in the high nineties. Those figures are generated overwhelmingly on a small number of benchmark datasets, principally the IBM HR Analytics set of roughly 1,470 records and a widely circulated Kaggle human resources set of roughly 15,000 records, and typically after synthetic class balancing through oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique). These are legitimate methodological choices for comparing algorithms against one another. They are not estimates of field performance on an employer's own workforce, and they should never appear in a business case as though they were. Practitioner assessment of realistic field accuracy for flight risk models is considerably lower, in the range of

seventy to eighty percent, and analysts in the field have expressed sustained skepticism toward vendor claims above ninety percent.

在这一应用场景中,百分之八十与接近百分之百之间的差距并非无关紧要的四舍五入误差,因为基础发生率很低。大多数组织的年自愿离职率远低于百分之二十,这意味着一个头条准确率数字很高的模型,其误判为将要离职的人数仍可能大大超过真正将要离职的人数。在离职模型中,假阳性指的是被错误地判定为可能离职的员工。

治理问题随之而来,而投资回报的框架掩盖了这一点。组织拿这个名字要做什么?如果答案是什么都不做,那么该模型没有产生任何回报。如果答案是任何形式的行动——无论是差异化的留任激励、调整后的发展投入、将其排除在长周期任务之外,还是改变继任考量——那么组织就已经基于一项预测,对某个个人做出了差异化的雇佣决定。这是一项雇佣行为,其合法性取决于是什么驱动了该评分。构建此类模型的指导意见通常建议纳入员工人口统计信息,年龄和司龄被列为为模型提供背景信息的输入项之一。一个将年龄输入模型、用输出结果分配发展投入、却无法重新还原某名员工为何得出该评分的雇主,已经构成了《雇佣年龄歧视法》上的风险敞口,却将其记入账本作为一项收益。

> **一种复合性运营风险**
>
> 评论人士还指出了一种任何商业论证中都没有列出的复合性运营风险:向未经培训、不知如何应对的经理传达离职风险评分,本身就可能促成该模型所预测的离职。

自动化合规并不能免除合规义务

关于人工智能能确保遵守劳动法及内部政策的主张应当从商业论证中彻底剔除,这并非因为这些工具毫无用处。能够标记漏休、加班阈值、分类异常或休假管理错误的监控系统确实可能具有价值,支持其运营方面的理由是合理的。

The word that fails is ensures. A monitoring system does not discharge a legal obligation; it produces outputs the employer then owns. If the system flags a potential violation and no one acts, the organization has created a dated internal record establishing notice, which is materially worse than not having monitored. If the system fails to flag a violation that occurred, the employer has acquired no defense, because reliance on a vendor tool is not a recognized excuse for non-compliance under the Fair Labor Standards Act or the statutes the Equal Employment

Opportunity Commission (EEOC) enforces. If the system's determinations are wrong in a patterned way, the employer has automated the error and scaled it.

2025年1月发生的事件使这一情况雪上加霜而非有所改善。当EEOC从其网站上撤下人工智能相关指导文件,包括2022年5月关于《美国残疾人法》的技术协助文件和2023年5月关于第七章的技术协助文件时,雇主失去了原本告知其哪些做法被该机构视为存在问题的解释性材料。相关法规本身并未改变。一个依据已被撤回的指导意见配置的合规工具,如今在运行时已没有明确的参照标准,而这一空缺是由雇主而非供应商来承担的。

因此,合规自动化增加的是文档记录和应对义务,而不是免除了这类义务。这项义务是有成本的,这一成本是真实存在的,理应列入模型的成本一侧。

模型所遗漏的成本

大多数人力资源人工智能商业论证的成本一侧,仅包括许可费、实施服务费,以及某种程度的培训预算。那只是软件的采购成本。它并非运营一套自动化雇佣决策系统的成本,后者至少还应包括以下内容。

验证工作,即确立每个系统所评分的内容与该职位有文档记录的要求相对应的分析工作,以及在系统或职位发生变化时所需的重新验证。便利措施路径的运行,即随时能够根据申请提供替代性评估的能力,包括指定的决策人员、明确的处理时限,以及一个真正运作的互动流程,而非一张表格。记录保存与重现能力,即在29 C.F.R. 第1602部分所规定的保存期限内,能够针对任何一项个人决定,重现其评分、输入数据以及当时生效的系统配置。监控,即对结果进行定期测试,而非依赖供应商提供的摘要。供应商在取证和行政调查中的配合,这一项成本高昂,除非另行协商,否则不包含在标准商业条款之中,而一旦需要,费用便由雇主承担。以及内部治理职能本身,即需要有一名明确对上述事项负责的人员,而在大多数组织中,这一职能目前并不存在,必须新设或外包。

伴随这些运营成本的,还有一项几乎没有任何商业论证会计入价格的或有责任,而这一责任已不再是假设性的。2023年8月,EEOC与iTutorGroup就一款被配置为拒绝超过特定年龄的申请人的软件达成和解,金额为365,000美元,这是一个简单配置决定所对应的具体数字。2024年7月,加州北区法院裁定,针对一家甄选软件供应商提起的诉讼可以基于该供应商作为使用其工具的雇主的代理人这一理论继续推进,这动摇了一种常见假设,即责任要么完全归雇主,要么完全归供应商,并且可以通过合同清晰地加以分配。2025年3月,美国公民自由联盟(ACLU)、科罗拉多州ACLU分会、Public Justice以及Eisenberg and Baum代表一名聋人及原住民员工,就一次自动化视频面试以及被拒绝的字幕便利措

施申请,向科罗拉多州公民权利处及EEOC提交了投诉,两家涉事公司均对此予以否认。该项主张是针对个人的,不需要任何比较性统计数据,提起成本低廉。

正确的处理方式并非猜测一个和解金额,而是明确陈述该风险敞口,指出系统性做法方面的索赔往往不在标准雇佣行为责任保险的承保范围之内,并要求将降低该风险敞口所需的治理管控措施纳入部署预算,而非推迟处理。

加上这些成本项目的组织常常会发现,这项投资仍然是合理的。这是一个不错的结果,但这并非重点所在。重点在于,由此得出的数字能够在董事会、审计师乃至最终的法庭面前站得住脚,而目前的数字做不到这一点。

一种站得住脚的结构

以下方法能够产生一份经得起审查、且现有人力资源与财务职能即可构建的商业论证。

1. 在部署之前而非之后建立基线:招聘周期、单次录用成本、每笔交易的行政成本、错误率,以及按细分群体划分的离职率,每一项均需在此前至少四个季度加以测量。
2. 将自动化带来的节省与人工智能带来的节省区分开来。说明哪些所宣称的收益源自基于规则的工作流处理,即便不采购人工智能也能获得,并单独为其定价。
3. 在测量录用质量的任何变化之前,先针对自身职位明确定义录用质量,说明所用的标准衡量指标,并说明其局限性。如果无法定义,就不要宣称在这方面有所改善。
4. Refuse benchmark-derived accuracy figures for any predictive model. Require field performance on the organization's own data, with the base rate stated and false positive volume expressed in headcount rather than percentage.
5. 对于任何产生个体层面评分的模型,在部署之前先记录该评分将引发何种后续行动,并将该行动作为一项雇佣决定而非一项分析输出加以审查。
6. 将上述运营义务作为经常性成本定价,而非一次性实施费用。
7. 将或有负债作为一个披露项目列出,并说明降低该负债的控制措施,无论是否指定具体金额。
8. 在采购时将验证证据、现场性能数据和可复现能力作为合同条款要求提供。保留任何拒绝提供的记录,这本身就是关于组织认知情况的信息。

结语

There is a real business case for artificial intelligence in human resources, and it is more modest and more defensible than the one currently in circulation. It rests on cycle time, transaction cost, and error reduction in administrative processes, all of which can be substantiated. It does not rest on quality of hire, retention, or predictive accuracy claims that the function's measurement practice cannot support and that carry liability the model does not price.

那些能够从这些系统中获得持久回报的组织,并非最早采用或采购最多的组织。而是那些在部署之前就进行了测量、能够说明每个系统的评分结果及其与岗位的相关性、并将治理机制视为拥有成本的一部分而非需要规避的开支的组织。依赖未经测量的收益和未定价的风险所得出的回报,并不是真正的回报。那只是一种预测,而当有人要求展示具体运算过程时,两者之间的差异便会显现出来。

参考文献

- 美国公民自由联盟、科罗拉多州公民自由联盟、公共正义组织以及Eisenberg & Baum LLP律师事务所。向科罗拉多州民权司及美国平等就业机会委员会提交的投诉,2025年3月19日。
- EEOC诉iTutorGroup公司案,案号1:22-cv-02565(纽约东区联邦地区法院),2023年8月和解令。
- Mobley诉Workday公司案,案号3:23-cv-00770(加利福尼亚北区联邦地区法院),2024年7月驳回动议裁定。
- Society for Human Resource Management. "The Dangers of Using Predictive Analytics to Gauge Employee Flight Risk." 2023.
- 美国平等就业机会委员会。《美国残疾人法案与利用软件、算法及人工智能评估求职者与雇员》,EEOC-NVTA-2022-2,2022年5月12日,以及《特定议题:评估第七章下就业选拔程序中软件、算法及人工智能所产生的不利影响》,2023年5月18日。两份文件均于2025年1月27日从该机构网站撤下。
- 《雇员选拔程序统一指南》，29 C.F.R. 第1607部分；记录保存要求，29 C.F.R. 第1602部分。

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

引用方式

Grizzle, H. M. (2025, April 4). 未定价风险不是回报:为何人力资源领域的大多数人工智能商业案例作为商业案例都是失败的. Novara Consulting Group. https://www.novaracg.com/zh/2025/04/04/unpriced-risk-is-not-return-why-most-artificial-intelligence-business-cases-in-human-resources-fail-as-business-cases/