NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2025/06/07/instruments-not-principles-sequencing-ai-readiness-in-human-resource-governance/.

工具而非原则:人力资源治理中人工智能就绪度的排序

Heather Grizzle · June 7, 2025

Most AI readiness models put the ethics layer third. By then the systems are in production, procurement has closed, and governance is retrospective.

目录

1. 姗姗来迟的就绪模型
2. 为何原则不是管控手段
3. 颠倒过来的就绪顺序
4. 以人才智能为例进行说明
5. 这对咨询问题意味着什么改变
6. 参考文献

姗姗来迟的就绪模型

大多数组织性AI就绪框架都存在一个损害其效力的结构性特征。它们把能力建设排在管控之前。数据基础首先建立,应用第二步部署,而伦理或信任层则在第三步引入,通常以指导原则、宗旨以及对公平和透明的承诺来表述。这种顺序看似直观,因为它映照了工作实际获得资助的方式,也是目前流行的大多数就绪模型所推荐的顺序。

It is also the sequence that produces the failures now visible in employment AI. By the time an ethics layer is designed, the systems it would govern are in production, procurement has closed, contractual terms are fixed, vendor selection is sunk, and the organization′s practical options have narrowed to documentation and disclosure. Governance introduced at step three is not governance. It is retrospective justification of decisions made at step two.

纠正之道并非把伦理阶段提前。而是要认识到,伦理准则并非当前局面所需要的工具,把它当作工具本身就是错误所在。

第三步才引入的治理并非治理。

为何原则不是管控手段

原则与工具之间的区别并非语义之争,治理文献对此已有足够细致的探讨,组织无需重新发现这一点。

Salamon′s analysis of the tools of government establishes that policy objectives are realized through instruments with identifiable properties: what activity they authorize, who administers them, what compliance they compel, and what they make observable. A stated commitment does none of these things. It expresses an intention without specifying an actor, a trigger, an evidentiary output, or a consequence for non-performance. Applied to artificial intelligence, the pattern has been documented directly. Jobin, Ienca, and Vayena′s survey of the global landscape of AI ethics guidelines found substantial convergence on high-level principles alongside marked divergence in how those principles were interpreted and what they were understood to require. Mittelstadt′s assessment reached the sharper conclusion that principles alone cannot guarantee ethical artificial intelligence, in the absence of the professional structures, accountability mechanisms, and enforcement pathways that make principles operative in other domains.

因此,一个以制定清晰伦理准则来应对就业AI风险的组织,实际上采用了证据基础最薄弱的干预手段。这并非因为公平与透明不是值得追求的目标,而是因为一项准则不产生任何具体成果,而组织实际面临的每一项义务都是通过具体成果来履行的。

以几乎每个就绪模型中都会出现的“算法运作透明度承诺”为例。按照所述条件,这一承诺是无法实现的。供应商不会披露模型架构,该类别中的大多数系统在承诺所暗示的意义上并不具有可解释性,而承诺算法透明度的组织已经承诺了它无法兑现、且必然会被追问的事情。可实现且更有用的承诺属于不同性质:即由组织可以指认的一方,按照既定周期,针对受保护群体产出结果的书面证据。这才是一种工具。它明确规定了谁在何时做什么,以及产生何种记录。

颠倒过来的就绪顺序

一个建立在工具而非原则之上的就绪模型,以不同的方式排列工作顺序,而这种重新排序是实质性的贡献,而非表面形式上的改变。

Coverage determination comes first. Before any system is evaluated on capability, the organization must be able to answer whether the system′s outputs are used to make, guide, or materially influence decisions about hiring, promotion, compensation, assignment, discipline, or continued employment. That formulation, and not the product category under which a system is marketed, determines what obligations attach across every regime currently governing employment AI. Coverage determination is an instrument because it produces a record, assigns an owner, and triggers a defined workflow. It is also cheap, and it is the only step that becomes more expensive the longer it is deferred.

数据治理随之而来,但需重新定义框架。就绪模型将数据目标描述为“单一真实来源”,这是一个数据质量目标。治理目标则不同且更为狭窄:组织必须能够确定某一具体推断所依据的数据是什么、产生于何时、来源如何。整洁、整合的数据支持更好的预测。有据可查的数据来源支持组织向监管者或原告说明一个人为何得到其所获结果的能力。只有后者才是一种治理能力,而组织完全可以在不具备后者的情况下拥有前者。

证据性要求应在采购阶段就已介入,而非事后补入。这正是组织尚能保有议价权的时点。偏见审计文档需明确方法论、所测试的群体类别以及样本人群。确认没有任何地理、教育或司龄衍生特征在实际上充当代理变量。明确规定通知和不利行动流程,以及由哪一方执行该流程。留存与导出条款须足以在适用追诉期内经受平台迁移的考验。这些条款在中标前均可协商,而中标后则无法再行协商,这正是主张这种顺序安排的全部理由所在。

Deployment carries monitoring, not affirmation. The commitment that matters post-implementation is outcome testing at a stated interval, because under a disparate impact standard an organization′s belief that its tool is fair is not responsive to the question of whether outcomes differ. Only measurement is responsive.

The record layer closes the loop. Attestation bound to individual, version, date, and authenticated pathway. Retention architected at creation. The organization′s ability to demonstrate what it did is the output of the whole sequence, and it is the only output an adverse party will ever see.

这一模型是一个循环,而非一架梯子,因为第一阶段会反复出现。系统会不断进入组织,常常是通过那些并不将其归类为就业决策技术的职能部门,因此覆盖范围认定是一种持续受监测的状态,而非一个已完成的项目。

以人才智能为例进行说明

目前流行的就绪模型通常将主动型人才智能——即对现有技能与未来需求进行映射、生成数据驱动的职业发展路径——置于第二阶段,排在治理层之前。这种排序颇具启示性,因为这项特定能力恰恰是整个就业AI类别中监管最为严格的能力之一,却几乎从未被如此认定。

A system that maps employee capability and generates career path recommendations is producing assessments that materially influence promotion and assignment. New York City's Local Law 144 reaches promotion decisions. The Illinois amendment to the Human Rights Act, effective January 2026, reaches promotion, discipline, and discharge, and imposes notice obligations attaching to use rather than to outcome. Annex III of the EU Artificial Intelligence Act enumerates promotion, task allocation, and performance evaluation by function rather than by product category. An organization that deploys talent intelligence at stage two and reaches its trust framework at stage three has deployed a regulated system and then written its principles about it.

这就是常规顺序所带来的实际代价,而这并非假设情形。这正是组织获得其从未评估过的义务的通常路径。

这对咨询问题意味着什么改变

组织通常提出的问题是应当采用哪些AI工具。而真正决定其风险敞口的问题是,它们已经采用的系统中,哪些属于具有实质性影响的决策系统,以及针对每一个系统,它们持有哪些证据。

这是两个不同的问题,而第二个问题无需做出技术决策即可立即回答,成本低廉,且能产出一份持久的证据文件。它同时也是使后续技术决策具有可辩护性、而非仅仅出于热情的那个问题。一个能够就每一个涉及就业决策的系统说明其功能、所附带的义务、现有证据以及下一次审查负责人的组织,已经建立了某种东西。一个拥有一套伦理AI原则的组织,只是写下了某种东西。

将人工智能整合进人力资源运营并非一个技术项目。这一点常规模型说对了。它是一个治理项目,而治理是由工具构成的,而非由承诺构成的。

参考文献

- 萨拉蒙(Salamon, L. M.)编,《政府的工具:新治理指南》。牛津大学出版社,2002年。
- Jobin, A., Ienca, M., and Vayena, E. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1 (2019): 389–399.
- Mittelstadt, B. "Principles alone cannot guarantee ethical AI." *Nature Machine Intelligence* 1 (2019): 501–507.
- 美国国家标准与技术研究院,《人工智能风险管理框架*(AI RMF 1.0)*》,2023年1月。
- ISO/IEC 42001:2023,《信息技术——人工智能——管理体系》。
- 《纽约市行政法典》第20-870条及以下(2021年第144号地方法律)。
- 伊利诺伊州众议院第3773号法案,第103届大会(2024年),修订《伊利诺伊州人权法》,自2026年1月1日起生效。
- 欧盟法规2024/1689,附件三(4)。
- Novara Consulting Group, "The Development Exception: Internal Mobility Platforms and the Perimeter of Employment AI Governance," May 2025.

* * *

免责声明。
Novara Consulting Group 为机构性人工智能部署与劳动力合规基础设施提供治理与采购咨询服务。本文仅作分析之用,不构成法律建议。

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, June 7). 工具而非原则:人力资源治理中人工智能就绪度的排序. Novara Consulting Group. https://www.novaracg.com/zh/2025/06/07/instruments-not-principles-sequencing-ai-readiness-in-human-resource-governance/