NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/.

当控制系统继承了它们本应治理的同一种不确定性时

Heather Grizzle · April 18, 2026

When AI guardrails are themselves probabilistic models, governance inherits the uncertainty it is meant to control. What real control requires.

AI治理目前更多是根据什么最容易可视化来描述和推销的,而不是根据什么最值得分析来构建的。

这场讨论被可见的产物所主导:政策仪表盘、分层控制图、标注清晰的护栏,以及一步步呈现执行流程,暗示在任何行动发生之前,请求都经过了严格的评估。这些呈现方式之所以给人以受控的印象,是因为它们类似于传统的系统监督。然而,它们掩盖了一个更为关键的问题:在这些层级之内,究竟是何种机制在真正执行评估。

在目前大多数架构中,一个AI系统会生成一个建议行动,该建议随后被送入一系列检查流程,最终由另一个模型来判断该行动是否应当继续执行。这一评估层可能因供应商或框架的不同而有不同的描述方式。它可能被定位为策略引擎、风险评分机制,或是辅以规则的置信度排序系统。尽管术语上存在这些差异,其底层结构却始终一致:一个概率模型被赋予评估另一个概率模型所生成输出是否恰当的任务。

从风险与可靠性的角度审视,这一设计选择带来了重大影响。概率系统就其定义而言,在相同条件下并不会产生确定性的、可重复的结果。它们的输出是由源自训练数据的统计模式所塑造的,而它们的评估同样受概率推断而非固定标准的影响。当这样一个系统被用来治理另一个同类型的系统时,治理功能并不会引入一种根本上不同的可靠性形式。相反,它只是在更高的抽象层级上重复了同样的不确定性特征。

从治理的角度来看,这形成了一种循环结构。旨在实施控制的系统,与其所监督的系统一样,受到相同的变异性、边缘情形不稳定性和情境敏感性的影响。因此,增加更多层级、检查或封装并不必然转化为有意义的风险降低。它可能增加复杂性,也可能营造出尽职调查的表象,但它并未以能够产生明确控制边界的方式,改变系统的底层行为。

将其与其他高风险领域中治理的运作方式相比较,这一区别会变得更加清晰。例如,在金融系统中,交易审批流程依赖于预先设定的规则、审计日志以及可追溯、可重建的人为制定标准。在安全关键的工程环境中,控制系统被设计为具有与其所监控系统不同的可靠性特性。治理机制在功能和行为上都被刻意区分开来,以便在出现不确定性或故障时,能够提供一个稳定的参照点。

它引入的是递归,而非控制。

当前的AI治理模式往往未能建立这种区分。相反,它们将概率性判断叠加在概率性生成之上,并将由此产生的结构当作一个控制系统来对待。这种方法假设,在同一可靠性类别内增加更多评估步骤就能产生治理成效。而实际上,它引入的是递归而非控制,因为每一层都依赖于同一类型的推理过程,并容易受到同类故障的影响。

其后果绝非仅仅停留在理论层面。当一个系统产生造成损害的结果时,随之而来的问题并不是决策过程中是否涉及多个层级。真正的问题在于,这些层级中是否有任何一个真正发挥了控制作用——也就是说,它是否以与其所治理系统不同程度的确定性、可追溯性和问责制在运作。如果每一层都共享同样的概率性基础,那么这个系统就缺乏一个明确的支点,无法据此锚定责任,也无法可靠地重建决策过程。

决定性的检验标准

治理的定义在于引入一种运作可靠性特征与被治理系统不同的机制。

由此分析得出的立场是明确的。治理并非由系统中控制措施的数量所定义,也并非由其架构的视觉复杂程度所定义。治理的定义在于引入一种运作可靠性特征与被治理系统不同的机制。若缺乏这种区分,增加的层级并不构成控制。它们只是将同样的不确定性扩展到了更多组件之中。

当一个控制层的运作方式与其本应规范的系统如出一辙时,该架构并未在真正意义上实现治理。它产生的是一种递归结构,其中不确定性评估着不确定性,而行动与监督之间的边界也随之瓦解。真正的风险正潜藏于这一瓦解之中,无论该系统表面上呈现得多么井然有序。

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. (2026, April 18). 当控制系统继承了它们本应治理的同一种不确定性时. Novara Consulting Group. https://www.novaracg.com/zh/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/