人工智能作为治理技术：制度逻辑与公共信任研究

牟多铎*

（马来西亚理工大学，马来西亚 柔佛州 士古来 81310）

摘　要：随着人工智能算法在公共管理与社会治理中的普及，技术参与决策与资源配置的形式正在发生结构性变化。人工智能具有高速计算、预测建模和模式识别等能力，这使其在城市管理、公共服务与风险控制等领域具有工具性优势。然而，人工智能作为治理技术的制度适配仍存在协调问题，涉及透明性、责任机制与合法性基础等议题。制度逻辑的变化不仅影响公共决策方式，也重塑公众对技术治理的信任基础。公共信任是人工智能技术在治理场景中得以稳定运行的核心条件，信任结构取决于制度、认知与社会心理等多重因素。因此，有必要从制度逻辑与公共信任的交互关系出发，分析人工智能参与治理的结构性条件，为其制度化应用提供理论解释。本研究通过文献分析与治理案例梳理，探讨人工智能介入治理过程中的制度逻辑转换与信任机制构建问题，旨在深化人工智能技术在公共领域的社会科学理解。

关键词：人工智能；治理技术；制度逻辑；公共信任；技术治理

1 引言

1.1 研究背景

近年来，人工智能在公共管理与社会治理领域快速扩展，其应用涉及政务服务、城市管理、司法辅助与公共资源配置等场景。人工智能通过算法建模、风险预测与多源数据分析参与治理过程，在提高效率与降低人力成本方面展示出明显优势。特别是在数据基础设施逐步完善、行政流程数字化与治理场景可量化的条件下，人工智能对治理链条的介入呈现出由辅助性技术向制度性技术演化的趋势，不仅改变公共部门的信息处理方式，也改变公共管理系统资源配置与风险治理的逻辑。

然而，人工智能作为治理技术的制度适配过程仍存在诸多不确定性，涉及权责划分、技术透明性与治理合法性等问题，引发公共管理与社会科学界的广泛讨论[1]。与传统治理技术不同，人工智能的预测性判断与模型式决策具有过程不可感知、逻辑不可观察与结果不可推演的特征，这使得技术嵌入治理体系的制度前提更加复杂。无论从政府管理的流程理性层面，还是从公共服务的公平正义层面来看，人工智能参与治理都要求制度结构作出相应调整，以实现技术治理与制度治理之间的协调。

1.2 问题提出

人工智能介入治理并非单纯的技术替代过程，而是制度逻辑与社会秩序再组织的过程。人工智能技术对决策程序、制度安排与权力结构的影响已开始改变公共治理运行方式，使治理方式从经验型与规则型逐步向预测型与模型型转变。预测型治理强调通过算法模型识别模式与计算趋势，以降低不确定性并提升治理效率，这与传统制度对规则可解释性、程序可监 督性与权力可问责性的要求形成制度张力。

与此同时，人工智能的制度化进程存在公共信任基础不足、认知差异显著与风险感知复杂等问题，这使得技术治理在合法性与可接受性方面面临挑战。公众对技术治理的风险认知不仅涉及治理结果的不确定性，也涉及技术过程的不透明性与决策依据的不可解释性。制度逻辑与公共信任之间的互动关系因此成为理解人工智能治理的重要分析范畴，既关系到技术参与治理的制度边界，也关系到技术能否获得持续而稳定的社会接受。

1.3 研究意义

作者简介：牟多铎（1991-），男，博士研究生，研究方向为人工智能、计算机视觉、增强现实等。
通讯作者：牟多铎

本研究从制度逻辑与公共信任的视角分析人工智能参与公共治理的制度化条件，探讨技术治理在制度结构、决策程序与社会心理层面发生的变化，试图为人工智能的治理应用提供社会科学层面的理论解释。制度逻辑维度强调制度授权、权责结构与合法性基础，公共信任维度强调心理可接受性、风险评估与社会认同机制。两者的结合有助于解释人工智能技术如何从治理工具演变为治理制度要素。

此外，本研究不仅有助于深化对技术治理机制的认识，也有助于为政府与公共机构制定人工智能治理政策提供学术支撑[2]。在技术治理不断扩展的背景下，制度设计与公众认知成为人工智能治理可持续发展的关键。研究人工智能治理的制度化进程可为公共管理领域理解技术变革提供分析框架，为相关风险与挑战的应对提供制度启示，并为未来技术治理研究提供理论基础与研究空间。

2 制度逻辑的分析框架

2.1 制度逻辑的概念界定

制度逻辑主要用于解释制度安排与规范体系如何影响组织行为与决策过程，其强调制度结构、规则与认知基础在治理运行中的作用。制度逻辑不仅涉及正式制度，如法律规定、政策体系与组织规则，也涉及非正式制度，如社会认同、价值取向与文化心理。制度逻辑在公共治理中体现为对行为边界与行为合理性的界定，为治理主体提供行动依据与合法性基础。

在人工智能介入治理的情境中，制度逻辑涉及法律规范、组织机制、技术规则与社会认知等层面，并直接关系到技术治理的可行性与合法性[3]。与传统治理方式相比，人工智能技术具有预测性与模型化的工具属性，这使得制度逻辑不仅需要在规则层面调整适配，还需要在认知层面重构行为解释框架。制度逻辑的核心任务在于解释技术介入后治理体系如何保持制度稳定性与合法性，同时实现效率提升与风险控制的平衡。

2.2 技术治理与制度嵌入

技术治理是指利用技术手段实现公共事务管理与社会调控的方式，其强调技术在治理链条中的信息处理、决策引导与行为规制功能。人工智能技术在治理场景中的扩展表明技术正由辅助性手段转变为制度性工具，技术成为治理权力配置与规则运行的重要组成部分。

人工智能治理的制度嵌入包括监督机制配置、决策流程调整与问责规则重构等环节。制度嵌入过程要求在制度结构中明确技术权力的赋予方式、执行方式与监督方式，从而确保技术介入治理时具备合法性与可追溯性。制度嵌入的有效程度取决于技术的可解释性、可监督性与可问责性，这些制度变量共同决定人工智能在治理中运行的稳定性[4]。当制度无法有效吸收技术带来的不确定性时，技术治理可能引发制度风险，导致公众质疑制度与技术的合法性与可靠性。因此，技术治理的制度嵌入过程实际上构成了制度化技术治理的关键步骤。

2.3 制度合法性与社会接受

人工智能在治理领域的制度合法性不仅依赖于法律授权与规制体系，还取决于公众对技术治理的接受程度与信任基础。法律授权为技术赋予制度性权力，规制体系为技术行为提供规范边界，而公众接受则为技术治理提供心理基础与社会支持。三者共同构成人工智能治理的合法性结构。

合法性基础主要包括程序正义与结果正义两类维度，前者涉及决策过程的透明性与责任结构，后者涉及治理效果的公平性与效率性。程序正义强调技术治理需具备可解释性与责任链条，结果正义则强调技术治理需具备实效性与公共利益取向。人工智能治理的稳定运行需在程序与结果之间形成制度平衡[5]。当程序正义不足时，即便结果正义表现良好也可能难以获得公众支持；同样，当结果正义缺失时，即便程序正义完备也可能无法形成可持续信任。因此，制度合法性与社会接受的动态关系构成了人工智能技术治理制度化的基础。

3 人工智能作为治理技术的制度适配

3.1 技术特征与制度需求

人工智能技术具有数据驱动、模型预测与算法决策等特征，在治理场景中能够实现风险识别、行为分类与资源配置优化。相较于传统基于经验判断或规则匹配的治理方式，人工智能通过对大

规模数据的持续分析与模式识别，能够在复杂情境中提供预测性支持，从而提高治理的前瞻性与响应效率。然而，人工智能的技术特征并不天然适配既有制度结构，其运行逻辑与制度运行逻辑之间存在显著差异。

制度运行依赖可监督性与责任边界，强调决策过程的可追溯性与责任主体的明确性，而人工智能系统则倾向于通过概率模型与黑箱式决策实现效率提升。这种基于统计相关性而非因果解释的决策方式，使制度难以直接识别技术判断的依据与责任归属。这种差异导致制度对技术提出透明性与解释性要求，同时也对治理体系提出适配性挑战[6]。在缺乏有效制度调节的情况下，人工智能的技术优势可能转化为制度风险，影响治理体系的稳定性与合法性。

3.2 决策程序的技术介入

人工智能介入公共决策意味着决策程序可能从规则导向转向模型导向。传统公共决策通常基于明确规则、程序审查与人工判断，而模型导向的决策程序则依赖算法对数据的综合计算与预测输出。在行政审批、城市治理与公共服务等领域，人工智能技术通过算法模型参与判断、分配与协商，从而影响决策流程的程序结构与权力配置方式。

在这一过程中，决策程序不仅面临技术效率提升的问题，也面临程序正义与制度问责的问题。算法模型在提高处理速度与覆盖范围的同时，可能削弱人工审查在解释、申辩与纠错环节中的作用。制度化运行要求决策技术具备明确的责任链条与行为可追溯机制，从而保障公共决策的合法性基础。只有在制度层面对技术介入的边界、条件与责任作出清晰规定，人工智能参与决策才能在效率与正义之间形成可持续平衡。

3.3 治理规则的再组织

人工智能参与治理过程中可能引发治理规则的再组织，表现为制度权限结构的调整与制度边界的重塑。随着技术逐步嵌入治理体系，算法不再仅作为辅助工具存在，而可能在特定环节中获得准执行权或建议权，从而影响治理结果的形成过程。这一变化要求制度对人机关系、技术权限与决策层级作出重新界定。

技术治理在规则层面可能涉及授予算法执行权限、重新划分人机协同任务以及重新构建监督机制。例如，哪些事务可以由算法自动处理，哪些环节必须保留人工判断，如何对技术行为进行审计与纠错，均需要通过制度规则加以明确。从制度角度看，这意味着治理体系需要在规则、程序与技术之间形成新的平衡机制，以确保技术介入不会破坏既有制度秩序[7]。治理规则的再组织不仅是技术调整过程，也是制度自我修复与制度创新的重要体现。

4 公共信任的构成与作用机制

4.1 公共信任的概念结构

公共信任通常指公众对制度、组织或政策执行者的信赖基础，其形成依赖制度稳定性、执行一致性与价值正当性等因素。在公共治理研究中，公共信任被视为制度运行的重要社会条件，是公共政策得以有效实施的重要支撑。在人工智能治理情境中，信任对象由政府与组织扩展至技术系统，公众不仅需要信任治理主体，也需要信任技术本身及其运行方式。

在这一情境下，公共信任呈现出多层次结构，包括制度性信任、认知性信任与情感性信任等层面。制度性信任来源于对制度规则、法律授权与组织权威的认可，认知性信任来源于对技术能力、运行逻辑与治理效果的理解，情感性信任则来源于长期经验、社会心理与价值认同。上述信任层次相互交织，并受到风险认知、责任归属与制度透明等因素的共同影响[8]。人工智能治理的信任结构因此表现为制度、技术与社会心理的综合产物。

4.2 信任的社会心理机制

公众对人工智能的信任形成过程与风险感知密切相关。与传统治理方式相比，人工智能参与决策会改变公众评估风险的方式，使风险从基于个人经验与直接判断转向基于算法预测与模型输出。在这一过程中，风险不再完全可感知或可解释，公众对技术的理解程度直接影响其信任判断。

公众需要理解技术介入的目的、规则与后果，以便形成对技术治理的基本预期。如果技术运行方式过于复杂或不可解释，公众难以判断决策合理性与责任归属，可能导致信任成本上升，从而影响技术治理的合法性基础。社会心理层面的不确定感、被动感与不可控感，会削弱公众对制

度与技术的信任。因此，信任机制不仅是技术问题，也是心理与认知问题，其形成过程与信息透明度、沟通机制与参与渠道密切相关。

4.3 信任与制度合法性

公共信任是人工智能治理得以制度化运行的重要条件。信任不仅关系到技术是否能够获得公众认可，也关系到制度合法性是否能够稳定维持。制度合法性需要通过制度规则、程序设计与治理效果不断被确认，而公共信任则为这一过程提供社会基础。人工智能治理若缺乏信任支持，即便在技术层面具备优势，也难以形成长期稳定的制度安排。

合法性基础越薄弱，技术治理越难获得社会支持，公众越可能对技术介入治理持怀疑态度；合法性基础越稳固，技术越可能成为制度化工具，被纳入常态化治理体系。因此，信任机制构成制度化人工智能治理的社会心理支撑，其作用在于连接制度权威与公众认知，缓冲技术不确定性带来的治理风险，并为技术治理的持续运行提供合法性基础。

5 人工智能治理场景中的信任机制

5.1 技术信任与制度信任的区分

在人工智能治理情境中，信任对象呈现出明显的多层结构。公众既需要信任技术系统本身的运行能力，也需要信任配置与监管技术权力的制度主体。技术信任主要围绕算法模型的准确性、系统运行的稳定性以及技术决策的可解释性展开，其核心在于技术是否能够在既定目标下持续产生可靠结果。制度信任则侧重于制度权威、规则正当性与责任机制，强调技术运行是否受到制度约束以及制度是否具备纠错与问责能力。

在人工智能参与公共事务时，技术信任与制度信任往往相互关联，但其形成逻辑存在显著差异。技术信任多源于操作性结果，公众通过技术表现与实际效果逐步形成判断；制度信任更多源自制度合法性与制度权威，其基础在于制度规则的稳定性、程序的正当性与责任安排的清晰性[9]。当技术信任缺乏制度支撑时，技术表现的短期优势难以转化为长期信任；当制度信任不足时，即便技术性能优越，也可能因公众质疑制度授权而受到抵制。

5.2 责任机制与可解释性

人工智能介入决策后，责任归属问题成为影响信任机制的重要因素。人工智能系统通常以概率模型与数据模式参与判断与预测，其决策过程往往难以通过传统方式进行解释，这可能造成责任主体与执行行为之间的脱节。治理制度若无法明确区分技术建议、自动执行与人工裁量之间的责任边界，公众将难以判断决策失误的责任归属，从而削弱对技术治理的信任。

制度规范要求公共决策行为具备可追溯性与可问责性，因此技术治理需要构建能够识别技术行为、分配责任的制度体系。这不仅包括对算法设计与数据来源的监管，也包括对技术应用后果的责任追究机制。此外，人工智能系统的可解释性直接影响公众对决策过程的理解程度。解释越充分，公众越容易形成合理预期，信任成本越低，制度接受度越高[10]。可解释性因此成为连接技术运行与公共信任的重要制度桥梁。

5.3 风险认知与心理预期

公众对人工智能治理的态度受到风险认知的显著影响。风险认知不仅涉及决策结果的不确定性，还涉及制度与技术之间协同运作的可靠性。人工智能技术与传统治理方式相比，其预测过程具有不可感知性与不可观察性，公众难以通过直观经验判断技术决策的合理性，这会改变公众对治理风险的心理预期。

在此情境下，公众对人工智能治理的信任往往取决于制度是否能够降低技术带来的不确定感。信任机制能否形成，取决于制度能否通过政策设计、信息公开与参与机制降低风险认知的心理成本，增强公众对制度运行的可理解性与可预测性。当制度能够提供稳定的信息反馈与纠错渠道时，公众更容易接受技术介入治理，从而形成相对稳固的信任关系。

5.4 情境依赖与信任建构

公共信任的形成具有显著的情境依赖性。不同治理领域的信任结构存在差异，例如司法辅助决策涉及公共权威与公正原则，城市治理强调服务效率与资源配置，公共卫生治理则关注风险控

制与集体安全。在不同治理场景中，公众对人工智能技术的信任关注点并不一致。

在这些领域中，人工智能技术能够获得的信任基础取决于技术介入的公共意义、制度边界与后果评估机制。如果技术介入被认为有助于公共利益，并且制度能够清晰界定技术权力的使用范围与责任边界，公众更容易形成积极信任。反之，当技术介入缺乏明确制度约束或公共解释不足时，信任机制难以稳定建立。因此，信任机制的构建必须在具体治理情境中进行分析，以便形成具有制度合理性的解释。

6 讨论

6.1 制度逻辑与公共信任的交互关系

人工智能介入公共治理表明制度逻辑与公共信任之间存在结构性关联。制度逻辑决定技术治理的制度边界、权力配置方式与运行规则，是技术得以进入公共领域并参与治理的前提条件；公共信任则决定技术治理能否获得公众认可，是制度安排得以稳定运行的重要社会基础。二者并非单向作用，而是在人工智能治理实践中形成相互塑造、相互强化的互动关系。

制度逻辑通过规则制定、程序设计与监督机制强化治理过程中的组织协调与制度稳定性，为人工智能技术的运行提供规范框架。公共信任则通过公众对制度稳定性、责任机制与决策后果的心理评估，对制度逻辑的有效性作出社会反馈。人工智能技术在制度逻辑层面强调规范性、合规性与可问责性，在信任层面强调可解释性与可接受性，二者共同影响技术治理的制度化进程。当制度逻辑能够回应公众对信任的基本期待时，技术治理更容易获得社会支持，反之则可能引发合法性争议。

6.2 制度化条件下的技术治理挑战

人工智能技术的制度化面临两类主要挑战。第一类为制度层面挑战，包括责任链条界定、法律规范适配、监督机制设计与治理边界划分等问题。随着人工智能逐步参与决策与执行环节，传统以人为核心的责任体系面临调整压力，制度需要重新界定技术在治理链条中的位置，以避免权责模糊与制度真空。

第二类为社会心理层面挑战，包括风险认知复杂化、信任成本提升与公众参与不足等问题。人工智能技术的运行逻辑对公众而言往往具有抽象性与不可感知性，这加剧了风险认知的不确定性。如果制度无法通过信息公开、解释机制与参与渠道缓解公众的不安感，技术治理可能在社会层面遭遇抵触。制度化条件要求在技术能力、制度权威与公众认知之间建立协调机制，从而确保人工智能在治理中的角色具备合法性与稳定性，并避免技术优势转化为制度风险。

6.3 多元主体的协同机制

人工智能治理涉及政府、技术机构与公众三类主体，其协同关系构成技术治理能否顺利运行的重要条件。政府作为制度设计者与公共权威，负责规则制定、制度授权与监督安排；技术机构作为技术实施主体，负责系统开发、运行维护与技术解释；公众则通过使用体验、风险评估与舆论反馈参与对技术治理的社会评价。

多元主体之间的协调关系决定技术治理能否形成制度闭环。如果政府能够明确制度边界，技术机构能够承担相应责任，公众能够获得参与与反馈渠道，技术治理更容易获得制度支持与社会认同。协同机制越成熟，人工智能技术越可能在制度层面获得稳定授权，在社会层面获得持续信任，从而形成制度合法性与社会信任相互支撑的治理格局。

6.4 中国语境下的研究空间

在中国情境中，人工智能治理具有显著的制度优势与治理实践空间。中国城市治理体系在数据基础设施建设、行政协调能力与技术应用推进方面具备较高整合度，为人工智能参与公共事务提供了现实条件。同时，政府在公共治理中的主导角色使制度设计与技术部署能够在较短周期内形成规模化实践。

未来研究可进一步关注人工智能在公共服务、司法辅助、城市管理与公共卫生治理等领域的制度化进程，从制度逻辑与公共信任视角深化人工智能治理的本土化解释。通过分析不同治理场景中的制度安排与信任机制，有助于揭示中国情境下技术治理的运行特点与潜在风险，为人工智

能治理的持续发展提供经验基础与理论启示。

7 结论

人工智能作为治理技术的出现，标志着公共治理方式正在经历制度化与技术化并行推进的深层转型。人工智能技术在工具层面通过数据处理、模型预测与智能分析提升治理效率，在制度层面通过重塑决策程序、规则配置与权力结构改变治理运行方式，在社会心理层面则通过影响公众对技术与制度的认知与评价，重构公共信任的形成基础。这种多层次影响表明，人工智能不再只是外在于制度的技术手段，而正在成为嵌入治理体系的重要制度要素。

本文的分析表明，制度逻辑与公共信任之间存在持续互动与相互塑造的关系。制度逻辑为人工智能参与治理提供制度边界、规范框架与合法性条件，决定技术权力的配置方式与运行规则；公共信任则为技术治理提供社会支持与心理基础，影响技术治理的可接受性与稳定性。二者的协调程度直接关系到人工智能治理能否实现从技术应用向制度化运行的转变。

从制度运行角度看，人工智能治理的制度化需要在程序正义与结果正义之间形成平衡，在技术效率与制度责任之间形成机制，在制度设计与社会认知之间形成协同。只有当技术运行具备可解释性、责任机制清晰且制度安排能够回应公众信任诉求时，人工智能才能在公共治理中发挥长期作用，而不至于引发合法性危机或信任流失。

未来的人工智能治理研究有必要进一步强调跨学科对话，推动公共管理、社会心理与技术研究之间的深度融合。通过加强对技术治理制度设计与公共信任机制之间关系的系统分析，可以为人工智能参与公共治理提供更具社会基础的理论解释与制度支撑，也为技术治理的可持续发展提供重要参考。

参考文献：

[1] 徐家琦. 人工智能技术参与治理的政策逻辑与公众接受性研究[J]. 科学学研究, 2023, 41(2): 215-224.
[2] 刘国强. 人工智能治理的制度逻辑与知识框架[J]. 情报杂志, 2022, 41(1): 3-11.
[3] 王海军. 制度逻辑视角下的数据治理机制研究[J]. 情报理论与实践, 2021, 44(8): 12-18.
[4] 陈涛. 技术治理与制度嵌入: 人工智能的公共应用情境分析[J]. 公共管理学报, 2022, 19(3): 45-56.
[5] 张宁. 人工智能治理的合法性基础：程序与结果的分析框架[J]. 华中科技大学学报(社会科学版), 2023, 37(5): 79-88.
[6] 李晓峰. 算法决策的制度逻辑与技术限制研究[J]. 科技进步与对策, 2021, 38(24): 52-59.
[7] 吴凯. 人工智能治理规则的制度化路径研究[J]. 公共行政评论, 2023, 16(2): 102-115.
[8] 孙晓菲. 公共信任的结构与形成机制研究[J]. 社会科学战线, 2020, (3): 145-153.
[9] 林志明. 技术信任与制度信任的双重逻辑研究[J]. 社会, 2022, 42(4): 125-138.
[10] 韩彤. 算法问责与公共治理的责任机制研究[J]. 南京社会科学, 2023, (6): 56-65.

Artificial Intelligence as a Governance Technology: Institutional Logic and Public Trust

MOU Duoduo*

(Universiti Teknologi Malaysia, Skudai, Johor 81310, Malaysia)

Abstract: With the widespread adoption of artificial intelligence algorithms in public administration and social governance, the modes through which technology participates in decision-making and resource allocation are undergoing structural transformation. Artificial intelligence offers instrumental advantages in areas such as urban management, public services, and risk control due to its capabilities in high-speed computation, predictive modeling, and pattern recognition. However, the institutional adaptation of artificial intelligence as a governance technology remains contested, particularly with regard to transparency, accountability mechanisms, and the foundations of legitimacy. Changes in institutional logic not only reshape modes of public decision-making but also reconfigure the basis of public trust in technology-enabled governance. Public trust constitutes a core condition for the stable operation of artificial intelligence within governance contexts, and its formation depends on institutional arrangements, cognitive understanding, and social-psychological factors. It is therefore necessary to examine the structural conditions under which artificial intelligence participates in governance by focusing on the interaction between institutional logic and public trust, so as to provide a theoretical account of its institutionalization. Drawing on literature analysis and a review of governance practices, this study explores shifts in institutional logic and the construction of trust mechanisms associated with the integration of artificial intelligence into governance processes, with the aim of advancing a social-scientific understanding of artificial intelligence in the public domain.

Keywords: Artificial intelligence; Governance technology; Institutional logic; Public trust; Technology governance