·科学技术与社会·

人机“信任”何以可能？
——从功能性依赖到“有意义的人类控制”

How is “Trust” Between Humans and Machines Possible?:
From Functional Dependency to “Meaningful Human Control”

王亮 /WANG Liang　　刘昱菲 /LIU Yufei

（西安交通大学马克思主义学院，陕西西安，710049）
(College of Marxism, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049)

摘　要：研究表明，人类对人工智能系统形成的所谓“信任”，在本质上是一种基于性能预期的功能性依赖关系。这种依赖在认知不透明与情感投射的作用下容易转化为虚假安全感，并在人工智能系统缺乏道德主体资格与责任链条碎片化的情境中进一步形成“责任鸿沟”。本文创新性地将“有意义的人类控制”理论引入人机信任分析，通过“追踪”机制确保系统行为能够响应人类价值与道德理由，并通过“追溯”机制重建清晰的责任关联结构。由此，人机信任不再被理解为单纯的功能性依赖，而是一种建立在清晰责任归属与人类控制基础上的规范性人机关系。

关键词：人机信任　虚假安全感　责任鸿沟　有意义的人类控制

Abstract: Human “trust” in AI is essentially a form of functional reliance based on performance expectations. Under epistemic opacity and emotional projection, such reliance may generate a false sense of security, which further creates a gap of responsibility in the absence of AI moral agency and fragmented responsibility chains. This paper introduces Meaningful Human Control (MHC) into the analysis of human-machine trust, which helps ensure the alignment of system behavior with human values and reconstructs a clear responsibility chain through a mechanism of “tracing”. Human-machine trust should therefore be understood not as mere functional reliance but as a normative relation grounded in human control and clear responsibility attribution.

Key Words: Human-machine trust; False sense of security; Responsibility gap; Meaningful human control

中图分类号：TP18; B842　DOI: 10.15994/j.1000-0763.2026.09.011　CSTR: 32281.14.jdn.2026.09.011

学界多以信任为核心探究人机交互理论，但当下人工智能信任研究存有概念混杂、理论分歧问题，无法有效阐释智能场景交互特征与伦理隐患。本文辨析人机信任本质、梳理相关伦理难题，引入可行理论框架，助力搭建可信、尽责、可持续的人机信任体系。

一、功能性依赖：人对智能机器信任的实质

基金项目：国家社会科学基金后期资助重点项目“社交机器人‘拟人化’伦理风险研究”（项目编号：23FZXA011）。
收稿日期：2025 年 12 月 2 日
作者简介：王　亮（1985–）男，湖北黄冈人，西安交通大学马克思主义学院副教授，研究方向为科技伦理。Email: wangg85@163.com
刘昱菲（2003–）女，湖北武汉人，西安交通大学马克思主义学院硕士研究生，研究方向为科技伦理。Email: liuyufei1123@stu.xjtu.edu.cn

在人机关系迅速扩展且“信任”一词被频繁使用的当下，一个关键问题是：人机信任究竟指向什么？是心理依赖、技术可靠性，还是某种被投射的拟社会期待？早在二十多年前，图梅拉（Maj Tuomela）与霍夫曼（Solveig Hoffmann）便试图为信任建立可形式化的结构，并提出信任可以被精确地还原为一组由“理由–信念–结论”构成的条件链，从而在计算机程序中被模拟。依据他们的分析，“信任”可在逻辑上区分为三个层级：其一，“理性社会规范性信任”，这是一种基于能力、动机与社会规范的一致性信任，带有明确的规范约束意味；其二，“预测性信任”，其特征是不依赖道德或规范承诺，而是基于主体对他者行为“发生可能性”的判断；其三，“预测性依赖”，其特征是完全不涉及信任，只保留依赖行为的功利结构，以收益大于风险作为唯一依据。[1]在此基础上，图梅拉与霍夫曼认为，无论是何种信任，都可以被程序模拟，因为它们都可以被表述为一系列清晰的条件句：如果某一类理由成立，则生成某一类信念；如果生成的信念满足某种结构性组合，则可判断为“理性社会规范性信任”“预测性信任”或“预测性依赖”。[1]例如，当主体A需要主体B完成行为x时，A的信任可这样形式化表达：

理由：主体A认为主体B有能力完成行为x（能力理由），且在当前合作关系中有社会规范表明B应该完成x（规范理由），并且B自愿接受这一规范（动机理由）；

信念：A因此形成信念：如果B既能做x，又应当做x，并且愿意遵守规范，那么B会完成x；

结论：基于这一信念，A对B产生“理性社会规范性信任”，相信B会履行相应行为。

换言之，在图梅拉与霍夫曼的语境中，“信任”并不依赖意识、自我理解或情感卷入，而是可以被重构为一种功能性结构，也即，它（1）不需诉诸心理状态，而是由一套可形式化的条件触发逻辑所支撑；（2）不依赖主观体验，而依赖对信息、规则与预期行为模式的操作性组合。

随后，塔德欧（Mariarosaria Taddeo）在梳理相关文献并分析经典信任理论的基础上，提出了“电子信任”（e-trust）的概念。[2]她指出，在现代数字环境中，人与技术系统之间可以形成一种依据信息透明度、功能可靠性和可验证证据而建立的稳定关系，这种关系并不依赖主体的情感、心理状态或道德规范，而以系统设计、行为记录、性能评估等客观机制为支撑。[2]事实上，这一结构更接近拜尔（Annette Baier）所讨论的依赖而非信任。拜尔明确区分了信任与依赖：我们可以依赖许多事物（如一个准确的时钟），但只有在认为对方对我们怀有某种“善意”时，我们才会说自己信任了某个对象。[3]从这一点来看，由于电子系统并不具备“善意”这一道德属性，它无法满足信任的本质特征，因此“电子信任”不过是以一个新术语指代传统意义上的依赖关系。

相较于前述理论的概念包装，艾斯（Charles M. Ess）则从哲学人类学视角对人机交互中的依赖与信任进行了阐释。他将人机交互中的信任关系清晰地定位为一种功能性依赖，即，人工智能系统“是无实体的他者，不仅缺乏人类自主性的情感、理性与道德意志，而且还缺乏信任他者所需的具身性”，我们与信息通信技术的关系，绝非道德主体间的“交流”，而仅仅是一种主体对客体的使用，“我可以选择和判断依赖这些手段，但不信任这些手段。”[4]艾斯进一步指出，将这种工具性依赖误读为信任，等同于在技术对象上施加拟人化或万物有灵的假设，这不仅掩盖了人机关系的工具本质，也可能诱发对技术能力的过度预期，从而忽视真正的人类责任承担。[4]因此，将人机交互中的信任关系界定为功能性依赖，既是对传统信任概念边界的哲学性澄清，也为现实伦理问题的辨识提供了起点。

二、功能性依赖透视下的人机信任伦理风险

1. 基于认知黑箱与情感操控的虚假安全感

虚假安全感指的是用户在对系统的实际能力边界和运行逻辑认识不清的情况下，产生的一种超越系统真实可靠性水平的认知，即一种

“不当的信任校准”。[5]用户对人工智能系统的依赖，最初建立在对系统过往稳定且成功表现的观察与体验之上，系统通过高频次的正确决策，逐步在用户心智中建立起一种可依赖的形象。然而，这看似可靠的依赖关系之下，潜藏着巨大的认知鸿沟，对于绝大多数非专业普通用户而言，现代人工智能系统，特别是基于深度学习的模型，是一个不折不扣的认知黑箱。伯勒尔（Jenna Burrell）指出，机器学习算法的不透明性并非偶然，而是其技术本质所固有的，她将其区分为三种类型：意图不透明、技术不透明以及社会不透明。[6]在这种多层次叠加的“黑箱”结构之中，至少潜藏着三方面的风险：其一，用户几乎无法穿透系统表层所呈现的性能去把握其内部的决策逻辑、能力边界与潜在的失败模式。当技术运行机制以不透明的方式被封装起来时，用户得以进行监督、质疑和批判的认知基础将被根本性削弱，系统的运作反而成为一种不可追问的“技术事实”。[7]其二，长期且稳定的优良性能不仅不能化解这种认知上的隔阂，反而会因其带来的效率与便利，弱化用户保持怀疑的能力，使人们将顺手和习惯误以为是对系统真实能力的全面理解。可靠表现越是常态化，越容易让用户产生一切尽在控制之中的心理错觉，从而忽视系统在能力边界附近、在非典型情境中以及在极端或对抗性输入下可能暴露的脆弱性。[8]与此同时，信任水平在缺乏校准机制的前提下往往会被无意识地推高，使用户在无意间进入“信任越界”的状态，即信任程度超出了系统实际可承担的功能范围。[9]其三，当上述心理机制与技术黑箱共同作用时，人类监控与警觉性会随之持续下降，形成一种因可靠性累积而不断加深的“自满化”倾向。[10]用户在面对系统无法处理的极端案例时，不仅难以及时识别风险信号，也往往无法迅速接管控制权。

此外，虚假安全感的另一根源在于功能性依赖关系模式被有意识地引入情感维度。通过拟人化设计对人类的信任心理机制进行模拟，诱导用户在不自觉中将情感反应投向机器，从而使依赖关系在情感维度上变得更加隐蔽。纳斯（Clifford Nass）与穆恩（Youngme Moon）通过一系列实验揭示，人类会无意识地将适用于人际交往的社会规则，如礼貌、性别刻板印象乃至互惠规范等等，套用于计算机程序。[11]基于这一人类认知特性，人工智能与机器人的设计被有意识地导向“情感操控”。正如萨林斯（John P. Sullins）所洞察的，设计者通过赋予机器人友好的人格化语音、拟人化的外观设计，以及共情式的交互反馈，刻意营造出一种“类主体”的假象。[12]当用户与一个表现出“关怀”与“理解”的机器长期互动时，他们用于处理人际关系的脑区与情感模式被激活，导致其无意识地将原本适用于道德主体的、包含善意预期的“信任”感，错误地投射到作为纯粹客体的机器之上，这种状态被一些学者界定为情感式“过度信任”，其本质是“伪信任”。[13]这种关系使用户在情感上更难对系统产生怀疑和批判，从而在关键时刻放松必要的警惕，甚至放弃监管职责。

2. 道德主体缺失的“责任鸿沟”

当我们将智能系统界定为功能性工具时，我们实际上在认知上预设了一种严格的主客体二元结构：人类被安置为唯一具备意向性、道德理解力与责任感的主体，而人工智能系统则被规定为仅负责执行与产出效能的客体。[14]韦尔迪奇奥（Mario Verdicchio）与佩里尼（Andrea Perin）指出，道德主体的核心在于具备自主性与感知能力，而人工智能即便在功能上表现出自主性，也无法理解其行为的道德意义，更不能像人类那样出于道德理由而行动。[15]在这种道德主体性高度不对称的依赖结构中，人工智能被制度性排除在可归责者之外，由此为“责任鸿沟”（responsibility gap）的出现提供了可能。所谓“责任鸿沟”，指的是在自主学习系统造成重大损害时，人类在道德与法律层面无法明确界定、追溯并承担责任的结构性困难。[16]根据马提亚斯（Andreas Matthias）的分析，这种困境并非源于暂时监管滞后，而是由于自主系统在运行过程中持续学习、适应并形成不可预测的行为模式，从而系统性削弱了人类设计者与操作者的控制力。[16]

此外，这一困境还根植于技术能力扩展所带来的结构性变化：随着人工智能系统的作用范围在空间和时间上不断延展，人类与系统之间的因果链条被拉长，在责任链条的起点，传统归责模式就难以明确锁定道德责任主体。人工智能系统的生命周期涵盖设计、开发、训练、部署、监管与使用等多个环节，涉及设计师、工程师、数据科学家、企业、用户和监管机构等众多参与者，功能性依赖逻辑将行动和决策权分散在庞大网络中，同时为责任规避提供制度化和心理化的支持。一方面，它诱发责任转移心理。当关键任务被委托给高度可靠的工具，而人类仅充当监督或辅助角色时，责任感和参与度会系统性下降。[17]另一方面，它导致责任碎片化。当有害结果由众多参与者共同形成，且每个参与者的贡献均不足以构成充分或必要条件时，责任不再呈现清晰链条，而是被切割为难以追溯的微小片段。[18]每个参与者都隐藏在功能性依赖链条的功能角色背后，将责任推向网络其他节点，甚至会出现“代理责任洗脱”（agency laundering）现象。[19]

三、基于“有意义的人类控制”的人机信任及其实践

通过前文分析可知，功能性依赖框架为我们揭示了现实人机交互中信任关系的本质及其潜在伦理风险。然而，这一框架主要是一种诊断性视角，它帮助我们理解当下人工智能仍被定位为工具性存在时，人机信任关系为何不可避免地暴露出虚假安全感与“责任鸿沟”等伦理问题。但随着人工智能技术不断迭代，人机交互将不可避免地迈向更高水平的自主性甚至“拟主体化”形态。[20]面对这种新的技术生态，我们必须进一步追问：在高度自主的人工代理日益深度参与决策与行动的未来，人类如何与其构建一种既能避免上述伦理风险，又能凸显人类主体价值与意义的信任关系？

因此，迫切需要一个能够超越功能性依赖并支撑未来人机交互模式的理论框架，使我们在赋予人工系统必要自主性的同时，仍能确保人类对系统行为保持终极控制，并牢固维持责任归属结构。基于此，本文引入“有意义的人类控制”（Meaningful Human Control, MHC）理论作为核心分析框架。该理论最初由桑托尼（Filippo Santonide Sio）与霍温（Jeroen van den Hoven）在自主武器系统的伦理争论中系统化提出，旨在确保高度自主系统的行为在伦理上仍能被人类引导、解释与追责。[21]其后，MHC理论被广泛应用于自动驾驶、智能医疗等领域，成为负责任人工智能的重要规范原则之一。[22]MHC的核心洞见在于重新界定“控制”这一概念：控制不意味着人类对人工系统进行实时、细粒度的操作，而是要求系统在行动层面必须与人类的道德意图和价值判断保持对齐，并且其行为及后果能够清晰地追溯至人类责任主体。桑托尼与霍温将“有意义的人类控制”的实现归纳为两个不可或缺的必要条件：其一为“追踪”（tracking），即系统行为必须能够响应特定情境中相关人类的道德理由；其二为“追溯”（tracing），即系统的行为、能力及其结果必须能够追溯至至少一名具备恰当道德与技术理解的人类行动者（如设计者、使用者或监管者）。[21]下文将进一步论证，正是基于人机交互中的“追踪”与“追溯”这两个条件，“有意义的人类控制”能够为系统性化解功能性依赖框架下暴露的伦理困境提供可行路径，并由此为更成熟、更稳固的人机信任关系奠定必要的伦理前提。

1.“追踪”对虚假安全感的伦理风险消解

“追踪”要求系统的行为能够响应特定情境中的相关道德理由。这一条件直接针对虚假安全感的生成根源，即用户与系统之间普遍存在的信息不对称、能力误认以及价值对齐缺失。“追踪”条件通过两个关键机制——（1）划定能力与价值边界以及（2）促进透明交互——从根本上矫正用户对系统的误判与过度预期，使原本建立在情绪投射与认知遮蔽上的“安全感”重新回到以理解、判断与审慎信赖为基础的伦理关系之中。[23]

首先分析关于能力与价值边界的划定机制。虚假安全感的一个核心来源在于系统能

力范围的不透明。在“追踪”条件的框架下，引入“道德操作设计域”（Moral Operational Design Domain Moral, ODD）能够有效回应这一问题。[23]“操作设计域”（ODD）原为自动驾驶领域的专业术语，聚焦于技术操作层面，而“道德操作设计域”则对其进行了拓展，使得系统的设计不仅需要定义其“可以”（can）在何种条件下运行，更必须明确其在道德上“应该”（ought to）或“不应该”（should not）在何种情境下运行。[23]其要求设计者在损害发生前就担负起预见和规避潜在风险的责任，通过提高设计主体的前瞻性的意识来约束与规范人工智能。[24]这种道德预设不仅从源头上限制了系统因价值偏离而引发意外后果的可能性，更通过划定清晰的能力与价值边界，为用户提供了稳定的行为预期，是系统获得社会许可的前提，[25]从根本上夯实了人机信任的伦理根基。同时，“道德操作设计域”所确立的边界，也发挥着关键的认知校准功能，当一个系统被明确告知“能够做什么”以及“不应在何种情境下运行”时，其可以促使用户的信任水平与系统的实际可靠性实现精准匹配，[26]实现所谓“校准信任”。[27]

此外，引发虚假安全感的另一根源在于情感诱导。当用户的判断依据从事实滑向感受时，信任便极易出现偏置，导致对系统的过度依赖与控制错置。为应对这一风险，可以建立“共享表征”（shared representations）这一更高层次的透明交互机制。[23]具体而言，“共享表征”通过一系列结构化的设计机制，将人机互动从情感幻象的表层体验重新锚定到可观察、可检验的理解基础之上。其一，通过“玻璃盒”与“生态界面”设计，使人工智能系统的决策路径、行动意图与能力边界变得透明可读；其二，“共享表征”要求系统具备动态更新能力，人工智能不仅需根据人类行为推断隐含意图，也要能通过口头或文本形式与用户显性沟通，使其对人类理由的理解保持情境一致；其三，通过将人类行为模型纳入“交互规划算法”，使人工智能能够以更稳定、可解释的方式把握人类偏好与行动模式，从而减少因表面情绪线索而产生的理解偏差；其四，通过价值对齐技术（如逆强化学习）推断并重建系统的目标结构，使人工智能的行为策略在规范上与人类意图保持一致，避免因错误目标追踪而放大用户对系统的情感性信赖或功能性误信。[23]

2.“追溯”对“责任鸿沟”的弥合

“追溯”条件要求，系统的行为及其可能产生的后果应能够追溯到至少一名在设计、部署或操作过程中具备适当道德与技术理解的相关人类行动者。[21]为了实现这种“追溯”能力，必须在系统设计与运行环节采取可操作的措施，使人类行动者的权责、能力与道德理解得到充分体现。在此，我们将从两个角度讨论弥合“责任鸿沟”的途径：一方面，通过“能力–权限”匹配确保在高度智能化环境中人类行动者的权责与实际控制能力相一致；另一方面，通过构建“双向链接”，将系统行为与人类的道德理解明确连接，实现责任的可追溯与可审查。

“责任鸿沟”中归责失灵风险的首要根源在于，高度自主性的人工智能系统作为行为主体与传统意义上的道德主体存在分离。[28]当系统的自主行为产生后果时，人类参与者往往既非直接执行者，也可能因缺乏实际控制力而无法正当承担责任。若继续沿用传统责任框架，则会出现两种不合理现象：一方面，系统行为引发损害时，责任可能落在无法干预的操作员或设计者身上；另一方面，真正能够影响系统的人类主体如果没有明确权责，则可能无法被追责，从而导致归责失灵。为应对这一问题，必须确保责任承载的对象具有实际可控性，并且这种可控性建立在能力与权限匹配的基础上。弗莱米什（Frank Flemisch）等人将“能权匹配”原则总结为，“预先分配责任的控制范围应小于或等于已授予权限且具备相应能力的控制范围；授予的控制权限应仅限于或小于现有能力所能涵盖的范围。”[29]这意味着责任的合法性与可执行性被绑定。进一步分析，这种匹配逻辑需要同时满足以下条件：（1）权限保障，即赋予人类最终控制权，以确保责任归属清晰；（2）能力保障，即确保主体具备理解系统状态、

把握任务边界并调整系统行为的能力。[23] 只有当权限与能力同时存在时，责任才能成为可执行的约束，而非理论上的假设。

然而，需要注意的是，“责任鸿沟”更深层的原因在于现代技术行动结构本身发生了变化。正如约纳斯（Hans Jonas）所指出的，现代技术将单一行为的因果链条在时间与空间上极度延展，使行动的后果呈现累积性、延时性与不可逆性，从而改变了行动、结果与责任之间的基本结构。[24] 在这种延展性的框架下，行为的效应被分散在多主体、多环节与跨系统的复杂链条之中，人类行动者在因果链条中的位置不再是集中、可辨识的“源点”，而逐渐退居为被系统结构吸收的背景条件。[30] 问题也由此产生：当因果链条被技术系统拉长时，人类的道德位置却没有随之在结构上同步扩展。然而，要解决这一问题，仅仅保障主体的控制能力与权限匹配并不足够。一个行动要能够归责于某人，必须满足两个前提：其意图或价值必须进入行动的生成结构；且行动的后果必须能在道德上指向他。这正是在“追溯”条件下“前向链接”（forward link）与“后向链接”（backward link）所要实现的功能。[23] 其一，“前向链接”确保系统行为在结构上承载人类的道德理由。通过记录并规范设计算法、选择阈值、确定目标函数、采纳数据等关键性决策环节，人类的价值判断被预嵌入系统的初始条件，使系统后续行为并非自发涌现，而是在人的理由所构建的框架内展开。因此，系统的每一步行动都拥有可辨识的道德出处，具备追溯到人的理由来源的可能，从而恢复其可归责性。[23] 其二，“后向链接”则旨在满足如下要求：“对于人机系统行为所产生的任何后果，通向该后果的人类决策与行为应当是可识别的”。[23] 在此意义上，“后向链接”并不要求对人工智能系统，尤其是深度学习与大模型内部推理过程，提供完全还原式的技术解释。本文所强调的“后向链接”，旨在确立人工智能作为一种“社会–技术”系统在现实治理语境中的行为可回溯性。具体而言，这一机制通过模型与数据的版本化标识、运行与部署状态的日志化记录，以及算法输出与具体使用情境的并行留痕，使人机系统行为在事后能够被重新嵌入制度背景、组织分工与责任结构之中加以评估，从而在规范意义上回答责任判断的关键问题：发生了什么，以及在何种关键环节中，是谁作出了具有道德与责任含义的决定？[23]

3.“追踪”和“追溯”机制在实践中的应用

为避免“有意义的人类控制”仅停留于抽象规范层面，本文进一步选取自动驾驶与医疗人工智能两个高度自主、强风险且已广泛部署的应用场景，考察“追踪/追溯”机制在现实系统中的具体判定方式及其实现路径。在自动驾驶系统中，“追踪”的实践判定，核心在于“自动驾驶控制系统”（Autonomous Driving Control System, ADCS）是否能够在其“运行设计域”内，持续将车辆行为对齐于相关人类行动者的规范性理由。[31] 具体而言，在感知与决策层面，通过将“安全优先”“风险最小化”“交通法规遵循”等规范性价值转化为优化目标或硬性约束并嵌入决策算法。[32] 这一做法可被理解为一种“价值敏感设计”。[33] 在运行与更新层面，“追踪”机制还可通过“显性远端更新”与“隐性近端更新”两种设计方案加以强化。[31] 前者“强调通过政策与设计进行的有意识系统改进”，如引入自动紧急制动、车道保持辅助、驾驶员监控与接管提示等功能；后者则表现为系统在受控范围内的自适应学习能力，使ADCS能够在反复驾驶经验中优化对人类操作意图、交通互动模式与风险信号的响应策略。[31] 相较之下，“追溯”的实践重点并不在于系统行为本身是否“合理”，而在于当系统产生潜在危险或实际损害时，相关控制权与责任是否能够清晰地回溯至具备相应技术、认知与道德能力的人类行动者。在设计阶段，自动驾驶伦理与安全研究普遍强调“原始设备制造商”（Original Equipment Manufacturer, OEM） 应建立可审计的设计决策链条，以避免责任在复杂技术架构中被结构性稀释。[34] 在制度与运行层面，“追溯”的落实还依赖于监管机构对系统边界与责任分配的明确规定。例如，通过强制事故数据记录、运行日志保存与事故报告制

度，确保在异常事件发生后，能够区分算法失效、系统设计缺陷与人类监督不足等不同责任来源。[35]

与主要围绕环境感知与行为决策展开的自动驾驶系统不同，医疗人工智能并非仅作为一种外在的决策工具介入实践，而是直接嵌入临床判断这一高度规范化的行动结构之中。[36]在“追踪”机制下，医疗AI的伦理关键并不在于要求系统提供类人化或事后合理化的解释，而在于通过技术性的可理解性设计，使系统输出能够被纳入医师的临床推理与道德判断过程之中。[37]在医疗AI的具体实践中，这一要求通常通过一系列工程化设计加以实现，例如，在模型输出层同步呈现“预测置信区间”“不确定性估计”、训练数据分布的局限性，以及明确的适用人群与使用前提说明。[38]此外，医疗AI场景下的“追溯”机制可通过“医疗算法审计”来实现，这一机制系统性地记录训练数据来源、模型版本迭代、部署环境配置与使用场景等关键信息，为AI系统决策建立完整的审计轨迹。[39]在技术层面，“医疗算法审计”通常通过一系列具体的技术操作来实现对模型开发与运行过程的“追溯”。为此，相关研究提出了三类机制：一是“模型文档”（DocML），通过将模型用途、训练数据与开发过程建立关联，实现设计决策与责任主体的可追溯；[40]二是临床AI的“结构化日志框架”（如MedLog），对模型调用过程中的输入、推理与输出进行“事件级记录”，使系统运行状态可回溯；[41]三是医疗AI“数据溯源与审计轨迹体系”，对数据来源、模型版本及推理条件进行系统记录，以支持合规审计与责任认定。[42]

结　论

人机信任的可能性既非自然生成，也非单纯依赖技术功能，而深植于人类对人工系统能力、价值与行为的理解、设计与把控之中。[43]通过功能性依赖框架的分析，我们能够精准地洞察传统人机信任的结构特征及其内在风险：一方面，由认知黑箱与情感操控所共同导致的虚假安全感；另一方面，由于人工系统道德主体的缺位及责任链条的碎片化而形成的“责任鸿沟”。为应对这一挑战，本文引入“有意义的人类控制”理论，提出以“追踪”和“追溯”为核心的双重机制。前者通过“道德操作设计域”与“共享表征”，确保系统行为能够响应人类道德理由，划定能力与价值边界，并促进透明交互，从而将盲目的功能性依赖转化为基于认知清晰、价值对齐和伦理理解的理性信任；后者通过“能权匹配”原则与“双向链接”，将系统行为及其结果与具备道德与技术理解的人类行动者明确关联，从而有效弥合责任鸿沟。在这一框架下，人类既可赋予系统必要的自主性，又能保持终极控制与责任归属，使人机信任从功能性依赖的脆弱假象转化为稳健、可持续且具有伦理正当性的关系。此外，随着人工智能自主性不断增强，人机协作的伦理挑战愈发复杂，这不仅要求技术设计者和监管者在“可控性”“合理性”及“重大相关问题”上承担前瞻性责任，也呼唤普通用户提升批判性素养，以应对AI驱动的社会发展所带来的伦理问题。[44]

［参 考 文 献］

[1] Tuomela, M., Hofmann, S. 'Simulating Rational Social Normative Trust, Predictive Trust, and Predictive Reliance between Agents'[J]. *Ethics and Information Technology*, 2003, 5(3): 163–176.

[2] Taddeo, M. 'Defining Trust and E-Trust: From Old Theories to New Problems'[J]. *International Journal of Technology and Human Interaction*, 2009, 5(2): 23–35.

[3] Baier, A. 'Trust and Antitrust'[J]. *Ethics*, 1986, 96(2): 234.

[4] Ess, C. M. 'Trust and Information and Communication Technologies'[A], Simon, J. (Ed.) *The Routledge Handbook of Trust and Philosophy*[C], New York: Routledge, 2020, 410.

[5] Lee, J. D., See, K. A. 'Trust in Automation: Designing for Appropriate Reliance'[J]. *Human Factors*, 2004, 46(1): 50–80.

[6] Burrell, J. 'How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms'[J]. *Big Data & Society*, 2016, 3(1): 1–12.

[7] Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*[M]. Cambridge: Harvard University Press, 2015.

[8] Wallach, W., Allen, C. *Moral Machines: Teaching Robots Right from Wrong*[M]. Oxford: Oxford University Press, 2009.

[9] Gebru, B., Zeleke, L., Blankson, D., et al. 'A Review on Human-Machine Trust Evaluation: Human-Centric and Machine-Centric Perspectives'[J]. *IEEE Transactions on Human-Machine Systems*, 2022, 52(5): 952–962.

[10] Parasuraman, R., Manzey, D. H. 'Complacency and Bias in Human Use of Automation: An Attentional Integration'[J]. *Human Factors*, 2010, 52(3): 381–410.

[11] Nass, C., Moon, Y. 'Machines and Mindlessness: Social Responses to Computers'[J]. *Journal of Social Issues*, 2000, 56(1): 81–103.

[12] Sullins, J. P. 'Trust in Robots'[A], Simon, J. (Ed.) *The Routledge Handbook of Trust and Philosophy*[C], New York: Routledge, 2020, 313–325.

[13] Hancock, P. A., Billings, D. R., Schaefer, K. E., et al. 'A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction'[J]. *Human Factors*, 2011, 53(5): 517–527.

[14] Himma, K. E. 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?'[J]. *Ethics and Information Technology*, 2009, 11(1): 19–29.

[15] Verdicchio, M., Perin, A. 'When Doctors and AI Interact: On Human Responsibility for Artificial Risks'[J]. *Philosophy & Technology*, 2022, 35(1): 11.

[16] Matthias, A. 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata'[J]. *Ethics and Information Technology*, 2004, 6(3): 175–183.

[17] Wickens, C. D., Helton, W. S., Hollands, J. G., et al. *Engineering Psychology and Human Performance*[M]. 5th Ed. New York: Routledge, 2022.

[18] Nissenbaum, H. 'Accountability in a Computerized Society'[J]. *Science and Engineering Ethics*, 1996, 2(1): 25–42.

[19] Rubel, A., Castro, C., Pham, A. 'Agency Laundering and Information Technologies'[J]. *Ethical Theory and Moral Practice*, 2019, 22(4): 1017–1041.

[20] Zeigler, B. P. 'High Autonomy Systems: Concepts and Models'[A], Zeigler, B. P., Rozenblit, J. (Eds.) *Proceedings [1990] AI, Simulation and Planning in High Autonomy Systems*[C], Los Alamitos: IEEE Computer Society Press, 1990, 2–7.

[21] Santoni de Sio, F., van den Hoven, J. 'Meaningful Human Control over Autonomous Systems: A Philosophical Account'[J]. *Frontiers in Robotics and AI*, 2018, 5: 15.

[22] Calvert, S. C. 'Principles and Framework for the Operationalisation of Meaningful Human Control Over Autonomous Systems'[J]. *Science and Engineering Ethics*, 2025, 31(5): 27.

[23] Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., et al. 'Meaningful Human Control: Actionable Properties for AI System Development'[J]. *AI and Ethics*, 2023, 3(1): 241–255.

[24] Jonas, H. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*[M]. Chicago: The University of Chicago Press, 1984.

[25] Burton, S., Habli, I., Lawton, T., et al. 'Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective'[J]. *Artificial Intelligence*, 2020, 279: 103201.

[26] Cummings, M. L. 'Automation Bias in Intelligent Time Critical Decision Support Systems'[A], Harris, D., Li, W. (Ed.) *Decision Making in Aviation*[C], London: Routledge, 2017, 289–294.

[27] Steinmetz, S. T., Naugle, A., Schutte, P., et al. 'The Trust Calibration Maturity Model for Characterizing and Communicating Trustworthiness of AI Systems'[J]. arXiv preprint, arXiv: 2503. 15511, 2025.

[28] Königs, P. 'Artificial Intelligence and Responsibility Gaps: What Is the Problem?'[J]. *Ethics and Information Technology*, 2022, 24(3): 36.

[29] Flemisch, F., Heesen, M., Hesse, T., et al. 'Towards a Dynamic Balance Between Humans and Automation: Authority, Ability, Responsibility and Control in Shared and Cooperative Control Situations'[J]. *Cognition, Technology & Work*, 2012, 14(1): 3–18.

[30] van der Loeff, A. S., Bassi, I., Kapila, S., et al. 'AI Ethics for Systemic Issues: A Structural Approach'[J]. arXiv preprint, arXiv: 1911. 03216, 2019.

[31] Calvert, S., Johnsen, S., George, A. 'Designing Automated Vehicle and Traffic Systems Towards Meaningful Human Control'[A], Mecacci, G., Amoroso, D., Siebert, L. C., et al. (Eds.) *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*[C], Cheltenham, UK: Edward Elgar Publishing, 2024, 167–186.

[32] Calvert, S. C., Mecacci, G. 'A Conceptual Control System Description of Cooperative and Automated Driving in Mixed Urban Traffic with Meaningful Human Control for Design and Evaluation'[J]. *IEEE Open Journal of Intelligent Transportation Systems*, 2020, 1: 147–158.

[33] 刘宝杰. 价值敏感设计方法探析[J]. 自然辩证法通讯，

2015, 37（2）: 94–98.

[34] Falco, G., Siegel, J. E. 'A Distributed “Black Box” Audit Trail Design Specification for Connected and Automated Vehicle Data and Software Assurance'[J]. *SAE International Journal of Transportation Cybersecurity and Privacy*, 2020, 3: 97–111.

[35] The U.S. Department of Transportation. 'Automated Driving Systems 2.0: A Vision for Safety'[EB/OL]. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf. 2017–09–06.

[36] Elgin, C. Y., Elgin, C. 'Ethical Implications of AI-Driven Clinical Decision Support Systems on Healthcare Resource Allocation: A Qualitative Study of Healthcare Professionals' Perspectives'[J]. *BMC Medical Ethics*, 2024, (25): 12.

[37] Sokol, K., Fackler, J., Vogt, J. E. 'Artificial Intelligence Should Genuinely Support Clinical Reasoning and Decision Making to Bridge the Translational Gap'[J]. *npj Digital Medicine*, 2025, (8): 1–11.

[38] Fan, X. 'Position Paper: Integrating Explainability and Uncertainty Estimation in Medical AI'[A], *2025 International Joint Conference on Neural Networks (IJCNN)*[C], Rome: IEEE, 2025, 1–8.

[39] Liu, X., Glocker, B., McCradden, M. M., et al. 'The Medical Algorithmic Audit'[J]. *The Lancet Digital Health*, 2022, 4(5): e384–e397.

[40] Bhat, A., Coursey, A., Hu, G., et al. 'Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability'[A], *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*[C], Hamburg: ACM, 2023, 1–17.

[41] Noori, A., Rodman, A., Karthikesalingam, A., et al. 'A Global Log for Medical AI'[J]. arXiv preprint, arXiv: 2510. 04033, 2025.

[42] Rawles, T. 'AI and LLM Data Provenance and Audit Trails for Healthcare Technology: Building Trust Through Transparency'[EB/OL]. https://www.onhealthcare.tech/p/ai-and-llm-data-provenance-and-audit. 2025–06–14.

[43] 董文莉、方卫宁. 自动化信任的研究综述与展望[J]. 自动化学报, 2021, 47（6）: 1183–1200.

[44] 陈小平. 人工智能伦理治理：一种新型问题的底层逻辑和创新探索[J]. 生命科学, 2022, 34（8）: 983–989.

[责任编辑　李斌]