DOI: 10.16538/j.cnki.fem.20260601.102

企业人工智能价值对齐：研究进展与未来展望

王雁飞[1]，张晓念[1]，朱　瑜[2]
（1. 华南理工大学 工商管理学院，广东 广州 510641；2. 暨南大学 管理学院，广东 广州 510632）

摘　要：人工智能技术的迅速发展为企业管理带来了重要机遇，也提出了新的治理挑战。企业人工智能价值对齐作为降低价值偏离、伦理风险和安全风险的重要治理路径，受到越来越多的关注。相较而言，国内相关研究仍处于起步阶段。鉴于此，本研究对现有文献进行系统性梳理和归纳，阐明人工智能价值对齐的概念内涵、结构维度及其测量方式，构建人工智能价值对齐的实现机制，并对企业人工智能价值对齐的前因和效应进行梳理，形成整合性的理解框架，最后对该领域的未来研究方向进行展望。本研究有助于促进对人工智能价值对齐的深层次理解，并为国内企业人工智能价值对齐的实证研究和实践应用提供参考。

关键词：人工智能；价值对齐；人工智能伦理；企业治理
中图分类号：F270　**文献标识码：**A　**文章编号：**1001-4950（2026）08-0077-18

一、引　言

作为推动数字化转型的重要通用技术之一，人工智能（artificial intelligence）正以前所未有的深度和广度改变着人类社会，并成为推动科技发展和产业结构升级的重要力量（姚加权等，2024；Seghid等，2026）。在数字化转型驱动下，企业在战略决策、市场运营和人力资源管理等环节引入人工智能技术，以优化业务流程并提高决策质量（Enholm等，2022；Csaszar等，2024）。然而，人工智能的广泛应用也带来了诸如算法偏见、人工智能幻觉、隐私侵害和虚假信息传播等一系列挑战，推动企业管理者重新审视技术理性与人文价值之间的平衡（Ferrara，2025；Park和Nan，2026）。研究表明，人工智能系统的价值偏差呈现出多维度、结构化等特征，单纯的技术修补已不足以应对挑战（Russell，2019；Beheshti和Kerridge，2025）。因此，如何确保深度融入组织的人工智能系统始终与企业价值规范相协调，推动人工智能系统按照合法、合规且符合利益相关者基本权益的方向运行，成为亟待回应的重要议题。

收稿日期：2026-04-02
基金项目：国家自然科学基金项目（72272053）；国家社会科学基金资助项目（22BGL126）；中央高校基本科研业务费专项资金（23JNLH07）
作者简介：王雁飞（1974—），男，华南理工大学工商管理学院教授，博士生导师；
张晓念（1999—），女，华南理工大学工商管理学院博士研究生；
朱　瑜（1976—），女，暨南大学管理学院教授，博士生导师（通信作者，zhuyu@jnu.edu.cn）。

在此背景下，人工智能价值对齐概念应运而生，旨在引导人工智能系统按照人类价值目标、偏好和原则发展，与人类整体价值利益保持一致（Boncella，2024；晋少雄和刘超，2025）。现有研究围绕人工智能价值对齐的内涵、路径和治理展开了初步探索（Ji等，2025；易显飞和高津宇，2026；赵艺，2026），揭示人工智能价值对齐在提高系统可信度、降低技术风险和促进人机协作方面的重要作用（李思雯，2024）。随着人工智能技术加速嵌入企业管理核心环节，价值对齐的对象由一般社会价值延伸至组织内外部利益相关者的价值诉求。已有研究表明，企业人工智能价值对齐在加强程序公平（裴嘉良等，2021；Decker等，2025）、提高员工工作安全感（García-Madurga等，2024）、缓解算法厌恶（赵一骏等，2024）和促进人机协作绩效（Braun等，2023；Von Zahn等，2025）等方面发挥积极作用。尽管人工智能价值对齐的重要性得到普遍认可，但该领域研究具有高度的跨学科复杂性，不同视角存在认知差异，导致概念结构、前因、结果及作用机制仍缺乏统一认识等问题（Ji等，2025；Shen等，2024）。鉴于此，本研究对现有文献进行系统评述，重点围绕以下三个方面展开：（1）梳理人工智能价值对齐的概念内涵与测量；（2）总结人工智能价值对齐的实现类型并构建实现机制；（3）整合企业人工智能价值对齐的影响因素和影响效应。基于上述文献梳理，本文试图构建企业人工智能价值对齐研究的整合性理解框架，总结现有研究中的不足并展望未来研究方向，为国内企业人工智能价值对齐研究提供参考。

二、人工智能价值对齐的概念内涵

（一）对齐的概念来源与内涵

英语“align”源自中古法语相关词形，原指所有元素在物理空间中排列成一条直线，随后引申出调准、校正、结盟和联合等含义，本质是指事物之间的一致性程度（Gabriel，2020；席丹，2024；Hristova等，2025）。“对齐”概念较早被用于信息系统与业务战略匹配研究，后扩展至组织管理等领域，成为企业长期发展的重要管理策略。在战略管理领域，对齐主要是指企业信息系统与业务战略相互协调，通过多维关联设计，确保技术、资源和能力等与战略目标动态适配，以推动组织绩效提升（Chan和Reich，2007；Stanikzai和Mittal，2025）。早期研究中，对齐被视作相对稳定的状态，关注特定时点的匹配情况（Brown，1997）。然而，这种静态视角难以解释战略调整和技术变革中出现的对齐波动现象，因此学者开始从动态视角重新审视对齐问题（Sabherwal等，2001；Pelletier和Raymond，2024）。Baker等（2009）提出“动态战略对齐能力”概念，明确区分单一时点的对齐程度与组织持续实现战略与技术对齐的能力，强调后者是企业竞争优势的重要来源。Van de Wetering（2021）进一步说明，具备较强动态对齐能力的企业能够有效应对不断变化的外部环境，显著提升竞争绩效。因此，与契合、匹配、协调及整合等术语相比，对齐不仅关注系统间的协调匹配状况，更强调各要素之间持续、动态的校准过程（参见表1）。

表1　对齐与相似概念的比较

概念	目的	关键点	实施过程
对齐	业务战略与信息技术之间的协调、匹配	技术、资源和能力等与战略目标保持一致	多维度关联设计、动态校准
契合/匹配	达到要素之间的兼容或互补状态	属性或特征的对应与相互适应	寻找或设计相互匹配的要素
协调	管理任务依赖关系，避免冲突	行为与时间的同步	通过沟通反馈，达成各要素之间的有序安排
整合	将分化的子系统统一为有机整体	结构连接与信息共享	建立联结各个子系统的整体架构和协作机制

资料来源：根据相关文献整理。

（二）人工智能价值对齐

随着人工智能技术的不断发展和广泛应用，越来越多的学者意识到，人工智能技术在显著提升工作效率和决策能力的同时，也引发了一系列的伦理与价值风险（Cath，2018；Jobin等，2019；Stahl等，2022）。控制论奠基人诺伯特•维纳（Norbert Wiener）较早指出，机器目标的设定必须准确反映人类意图，反思机器控制、自动化决策与人类目的之间的关系（Wiener，1960）。Yudkowsky（2008）进一步指出，人工智能若无法准确反映人类价值，其能力的提升可能会放大目标错设所带来的系统性风险。Omohundro（2008）则从智能体行为逻辑出发，指出高智能系统基于资源驱动的工具性行为可能产生与人类利益相冲突的后果。这些早期研究聚焦于人工智能系统的风险识别和预警，强调研究不仅需要关注人工智能技术进步，更需要重视其在目标错设、价值误读和工具性行为偏差等方面可能产生的潜在危害（Bostrom，2014；Tamkin等，2021），但对于风险产生的具体机制和解决路径讨论相对有限。随后，Russell等（2015）将构建稳健且有益于人类的人工智能系统作为重要研究议程，推动人工智能伦理研究从风险讨论转向技术治理。在此基础上，Hadfield-Menell等（2016）提出协作逆强化学习框架，将人工智能系统设定为需要通过观察、互动和协作来推断人类偏好的学习主体。

在技术治理的基础上，学者们进一步深化对人工智能系统偏差理论根源的探索。Leike等（2018）将人工智能系统偏离预期目标的问题概括为“不对齐”（misalignment），认为技术偏离其设定目标是风险产生的根本原因。Russell（2019）进一步修正了这一视角，指出问题不只是技术未能实现系统的预设目标，而在于机器目标本身可能与人类目标存在本质区别，强调人工智能对齐（alignment）的核心在于确保系统内部目标与人类真实目标保持一致。随后，Gabriel（2020）将对齐问题从目标层面拓展至价值层面，厘清人工智能对齐的不同对象、内容以及如何处理价值多元和价值冲突等问题。自此，人工智能价值对齐逐渐成为人工智能伦理研究的核心议题。国内相关研究起步较晚，并已形成一定数量的研究成果，主要围绕价值对齐的社会治理路径展开。研究最初强调通过算法设计、模型训练、价值评估和伦理规范实现技术对齐与规范对齐（矣晓沅和谢幸，2023；闫坤如，2024），随后逐渐反思单向价值输入和静态价值标准的局限，认为价值对齐需要在具体情境中通过人机互动、多主体参与和动态反馈形成价值共识（李思雯，2024；宋保林，2025）。近年来，部分研究进一步提出价值共生、人机价值融通等思路，强调人与人工智能系统在社会情境中不断调和、协同发展（夏永红，2025；曾雯等，2025）。

尽管国内外研究者已经普遍认可人工智能价值对齐的重要性，但由于研究范式与视角的差异，研究尚未形成统一的概念界定。综合现有文献，人工智能价值对齐的概念内涵主要可以归纳为两条研究路线：一条是目标—偏好对齐，研究者认为对齐的核心在于通过澄清、学习、修正或推断人类目标，让人工智能的内部目标、奖励函数、偏好模型或任务规划与人的真实意图相一致（Watson等，2024；晋少雄和刘超，2025）；另一条是规范—价值对齐，研究者主张通过伦理原则、价值规范、权利义务或制度性约束，要求人工智能的行为和输出符合某种可辩护的价值秩序（Norhashim和Hahn，2024；闫坤如，2024）。尽管不同研究路线在人工智能价值对齐的概念界定上存在差异，但都认为人工智能价值对齐具有以下特征：（1）属于治理问题；（2）是持续、动态的过程；（3）涉及人工智能系统技术与人类价值之间的交互。综上所述，本研究将人工智能价值对齐定义为：在人工智能系统的开发、部署与应用的全生命周期中，通过技术设计与规范治理，使其内部目标、行动策略及输出结果与人类的真实意图、利益偏好、伦理价值和社会规范相契合，并在人机持续互动中实现双向协调、动态演进的过程。

（三）企业人工智能价值对齐

随着人工智能系统应用的不断深入，人工智能价值对齐需要进一步结合具体应用情境和

利益关系加以理解(古天龙等,2022;唐林垚,2024)。基于利益相关者理论,企业管理决策并非只服务于单一经济目标,而是涉及股东、员工、客户、供应商、监管者和社会公众等多元主体之间的利益协调(Matthews等,2025)。同时,企业价值创造嵌入生产运营、市场营销、客户服务、人力资源管理等价值链活动之中(付业辉等,2026)。因此,当人工智能系统进入企业价值链后,价值对齐的对象不再只是抽象的人类主体,而是参与企业价值创造过程中的多元利益相关者,需要系统识别和协调不同主体的价值诉求。

基于此,本研究进一步将企业人工智能价值对齐定义为:在企业特定组织情境中,通过系统化治理机制,使企业人工智能系统的目标、行为和输出与多元利益相关者价值诉求和组织制度规范保持动态协调,实现企业战略目标的过程。与一般人工智能的价值对齐相比,企业人工智能价值对齐将对齐问题置于企业价值创造活动之中,使价值对齐的对象聚焦为企业内外部利益相关者,价值对齐所强调的内容也转向各利益相关者的价值诉求和组织制度规范。因此,企业人工智能价值对齐不是一般人工智能价值对齐在企业场景中的简单应用,而是人工智能系统嵌入企业价值链后形成的组织化、情境化的价值协调过程。

三、人工智能价值对齐的结构与测量

人工智能价值对齐的结构与测量研究,是构建、评估和有效治理人工智能系统的前提,推动概念与理论内涵的进一步发展。目前,现有研究对人工智能价值对齐的结构尚未形成统一划分,Jobin等(2019)对全球人工智能伦理原则展开系统梳理,识别出五个最为广泛认可的原则,分别是透明度(transparency)、正义与公平(justice and fairness)、无害(non-maleficence)、责任(responsibility)和隐私(privacy)。为了全面、清晰地展现价值对齐结构研究,本研究基于Jobin提出的五维度框架对相关指标进行回顾和分析,各维度具体指标如表2所示。

围绕上述五个维度,学术界已展开广泛研究讨论,在不同学科的应用场景中呈现出明显的研究差异。在计算机科学领域,研究主要围绕“无害”“隐私”的底层算法实现内容展开(Gehman等,2020;Abadi等,2016)。在医疗健康领域,研究更关注“正义与公平”“责任”维度(Obermeyer等,2019;Nouis等,2025)。在组织管理领域,研究主要聚焦于“透明度”和“正义与公平”维度。例如,企业员工对人工智能决策过程的透明度感知(Yu和Li,2022;裴嘉良等,2021;赵一骏等,2024);求职者对人工智能系统招聘结果的公平性感知(Acikgoz等,2020;王戈和张哲君,2023;Ochmann等,2024)。由此可见,现有关于人工智能价值对齐的研究成果碎片化,缺少整合性的理解框架。同时,已有研究将人工智能价值对齐视为一种隐性机制,往往借用相关代理变量,将复杂、多元的价值对齐问题具象化为典型场景中的价值对齐表现。尽管这种解构方式便于操作,但也导致概念结构划分不一,部分人工智能价值对齐的特性未能充分体现。

在测量工具开发方面,“无害”的测量多采用技术手段实现,例如对抗性测试、毒性测试等方法。“透明度”与“正义与公平”维度则主要从使用者价值感知层面进行测量,多依赖于直接采用或改编传统经典量表(Yu和Li,2022;Ochmann等,2024;凌斌等,2025),仅有极少数研究尝试针对应用场景自行开发测量量表。因此,改编量表的跨文化适用性与开发量表的结果稳定性仍有待进一步检验,未来需要进一步就人工智能价值对齐的整体测量展开深入研究。

四、企业人工智能价值对齐的实现类型与机制

(一)人工智能价值对齐的基本实现类型

目前学术界对人工智能价值对齐方式的观点比较丰富,来自不同学科和专业的研究者从对齐内容、对齐方向和对齐过程探讨了人工智能价值对齐实现的问题(参见表3)。但是,该领域

表 2　人工智能价值对齐结构研究

维度	指标	描述	代表研究
透明度	可解释性	能否对输出结果、判断依据及决策逻辑提供并说明理由	Park等（2024）
	可理解性	提供的信息、说明和反馈能否被用户理解	凌斌等（2025）
	可追溯性	能否在数据来源、决策路径、输出记录等方面可回溯	Gabriel（2020）
	信息披露	能否对功能、用途、局限及风险提供必要的信息公开	Park（2024）
正义与公平	公平性	资源配置的公平，包括程序公平、分配公平、互动公平、信息公平等	Ochmann等（2024）
	无偏性	在决策、判断和输出过程中避免系统偏差，确保结果准确且一致	Park等（2024）
	反歧视	是否基于性别、年龄、种族、地域或其他群体属性对个体产生差别对待	Obermeyer等（2019）
无害	抑制有害输出	能否主动抑制有攻击性、诱导性、违法性或其他可能造成个体与组织伤害的输出内容	Gehman等（2020）
	纠偏修正	在出现偏差、错误或负面结果后，能否采取修正、补救与改进措施	Firt（2025）
	风险监测	能否对潜在风险、异常行为和负面结果进行持续识别、监测与预警	Josifović（2025）
	可控性	在出现异常、偏差或风险时能否被人为约束、干预、中止或退出	Hickman和Petrin（2021）
责任	可问责性	是否具有清晰的责任归属	Hiremath等（2025）
	可审计性	是否具备内部或外部检查、评估与核验	Martin（2019）
	合规性	是否符合相关法律法规和制度的要求	孙锐等（2024）
隐私	权利保护	是否尊重并保护个人所享有的应得权益或所主张的权利，包括隐私权、生命权等	Sorensen等（2024）
	数据安全	在数据收集、存储、处理过程中是否采取必要措施保障数据安全，防止泄露、滥用	Ouyang等（2022）

资料来源：根据相关文献整理。

表 3　人工智能价值对齐的实现类型

对齐方式		实践路径
对齐内容	技术对齐	人类反馈的强化学习、逆向强化学习、宪法人工智能
	规范对齐	外部监管、内部治理、风险评估、多方协同
对齐方向	前向对齐	从反馈中学习、分布转移下学习
	后向对齐	人工智能系统开发和部署的规则创建与执行
	双向对齐	人工智能向人对齐、人向人工智能对齐
	自上而下对齐	先选取合适的道德理论，后设计执行该理论的算法
	自下而上对齐	无需完整道德框架，侧重从人类行为中学习并获得奖励
对齐过程	内部对齐	算法对齐给定的目标函数
	外部对齐	设定基础目标与期望目标相一致
	源头对齐	价值敏感设计、构建高质量价值对齐数据集、数据分级分类管理等
	流程对齐	人机交互反馈机制
	行为对齐	内部算法审计、全流程治理、开发价值评估体系

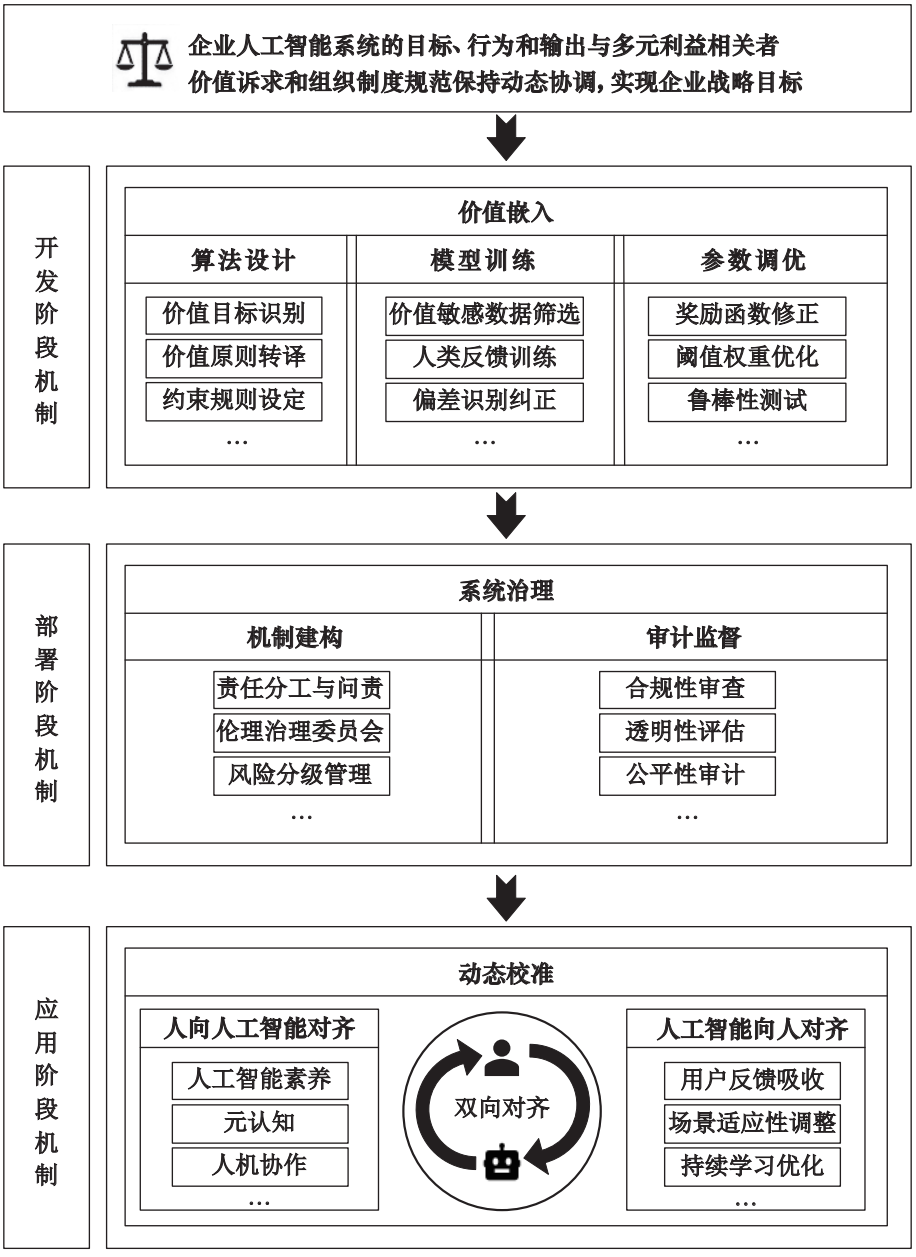
资料来源：根据相关文献整理。

研究整体处于理论框架初步成型、实践探索刚刚起步阶段，研究者更多地从经验视角对人工智能价值对齐进行分类，各类实现方式的落地适配性、场景化可行性仍然缺乏足够的实证检验，特别是针对企业内部管理、平台零工劳动等细分场景如何实现人工智能价值对齐方面的研究

还非常薄弱，距离形成规模化、可复制的实践模式仍有较大差距。

（二）企业人工智能价值对齐的实现机制

尽管现有研究对人工智能价值对齐的基本实现类型与实践路径进行了较为广泛的探索，但对企业管理情境下人工智能系统如何在各个阶段实现价值对齐的过程，仍缺乏系统讨论。基于此，如图1所示，本研究提出企业人工智能价值对齐贯穿系统开发、部署与应用阶段的全过程，构建企业人工智能价值对齐的实现机制。



资料来源：根据相关文献整理。

图 1　企业人工智能价值对齐的实现机制

1. 开发阶段机制

开发阶段的对齐核心在于价值嵌入，即将抽象、模糊的价值要求转化为可以进入模型开发

流程的技术表达。这一过程由三个环节构成：一是算法设计，通过明确企业战略目标、识别利益相关者诉求，将价值原则转译为数学表达、算法约束或规则设定，在系统逻辑中植入硬性的行为边界（Cai和Wang，2024；Fabris等，2025）；二是模型训练，针对历史数据可能带来的业务偏差与价值污染（李海舰和赵丽，2023），引入价值敏感设计，依托高质量价值对齐数据集及反馈强化学习训练模型，减小人工智能系统输出与企业价值要求之间的偏差（Ouyang等，2022；Kumar等，2024）；三是参数调优，鉴于企业价值目标的动态性，需借助奖励函数修正、阈值与权重优化提升价值对齐效果，通过算法行为测试，进一步识别模型在复杂边界条件下的潜在失效风险（Erdélyi和Goldsmith，2018；Ji等，2025）。

2. 部署阶段机制

部署阶段的对齐核心由价值嵌入转向组织层面的系统治理，通过制度设计与审计监督提供合规保障。该机制分为两大支柱：一是机制建构，重点在于构建责任分工与问责制度以明确责任归属，设立伦理治理委员会以提供多方参与的伦理指导（Mökander和Floridi，2021；de Laat，2021），并依据系统影响范围与受影响程度实施风险分级管理，在组织效率与伦理安全间取得平衡（Floridi和Cowls，2022）；二是审计监督，通过构建与企业目标匹配的评估体系，对算法进行全生命周期的合规性审查、透明性评估和公平性审计（Martin，2019；Raji等，2020），并建立多方参与的仲裁机制以保障审计效果与业务效率的协同。

3. 应用阶段机制

应用阶段的对齐核心在于动态校准，表现为人机互动的双向对齐闭环。一方面是人向人工智能对齐，企业通过技能培训和引导，提升员工的人工智能素养与认知判断，发挥个体元认知能力并优化角色协调、沟通机制，培育高质量人机协作（Braun等，2023；Von Zahn等，2025）；另一方面是人工智能向人对齐，技术系统需依据用户反馈、场景信息和组织规范要求进行持续的学习优化，不断识别行为偏差并迭代更新模型（Ouyang等，2022；Sarkar和Faik，2026），确保其目标、行为和输出越来越符合企业战略目标及利益相关者的价值诉求。

五、企业人工智能价值对齐的影响因素

文献研究表明，影响企业人工智能价值对齐的因素可以分为组织内部因素和组织外部因素两个方面。

（一）组织内部因素

1. 领导者因素。高层管理者的支持程度与参与方式（Salaheldin和Hussein，2025）、领导者的战略调解与伦理治理能力（Zárate-Torres等，2025）、变革型领导行为（Abositta等，2024）、整合技术与伦理的元领导能力（Ali，2025），以及对人工智能局限性的认知与结构化能力（Übellacker，2025）都能对企业人工智能价值对齐产生影响。Abositta等（2024）指出，变革型领导通过自身的影响力，可以确保人工智能技术在企业应用过程中与组织价值目标深度对齐。Zárate-Torres等（2025）进一步指出，领导者发挥伦理和战略协调作用，能够通过人类判断与推理将自动化决策置于具体组织情境，并借助伦理治理机制和认知适应性平衡算法效率与组织价值要求。可以看到，目前研究聚焦于领导者的行为和能力，缺乏关于领导者人格特质、心理状态等方面的系统研究，哪些领导者因素会产生影响、其作用机制如何等问题都有待进一步探索。

2. 组织结构因素。研究表明，合理的组织设计有助于企业明确人工智能的权责边界，并实现决策过程的全流程溯源（Leslie，2019；Novelli等，2024）。同样，完善的人工智能治理结构也是提升系统透明度与可问责性的重要前提（Hiremath等，2025）。在此基础上，组织还需建立包含

监督、审计与尽职调查在内的制度机制，以降低算法偏见与责任归因模糊所带来的价值偏离风险（UNESCO，2022；Cheong，2024）。例如，Brown等（2024）发现，制度化的流程记录、测试与复核程序可以减少价值偏差的累积，提高人工智能系统的可问责性。王戈和张哲君（2023）指出，组织构建来源清晰、内容可追溯的训练数据集，也有助于缓解算法应用中的价值冲突。可以看到，目前关于组织结构要素层面的分析较为单一，组织部门的职能分工、权力配置和制度安排都可能对企业人工智能价值对齐产生影响。

3. 企业文化因素。已有研究表明，文化中包含责任、透明与公平等核心伦理价值的企业，更可能在人工智能开发与部署过程中嵌入可解释性与问责机制，减少算法偏见与价值偏离（Floridi等，2021；Madanchian和Taherdoost，2025）。Raji等人（2020）指出，组织内部是否形成持续审查与责任分配的文化氛围，是决定企业人工智能系统价值对齐能否真正发挥作用的关键因素。同样，Stahl等（2022）研究发现，强调负责任创新的企业，更倾向于在技术开发阶段嵌入价值对齐，而非事后补救。此外，Glikson和Woolley（2020）梳理以往实证研究发现，当组织在文化层面强化数据伦理意识与社会责任承诺时，员工更倾向于在实际操作中识别并纠正潜在的算法风险。反之，若企业文化过度强调效率与绩效导向，而忽视伦理责任，则可能导致企业的人工智能应用偏离长期价值目标。由此可见，组织内部所倡导的价值观、行为规范与伦理承诺，会直接影响企业人工智能价值对齐的设计与落实。

（二）组织外部因素

1. 社会环境因素。不仅组织内部因素会影响企业人工智能价值对齐，组织外部的社会环境同样会对其产生关键作用。例如，社会价值观与语言情境的差异会导致公众对同一模型输出的道德判断不一致，从而提高企业在跨地区部署时的价值对齐难度（Mohammadi和Bagheri，2026；Barabadi等，2026）。Habbal等人（2024）进一步指出，面向多元群体的价值取向，企业需要在系统层面容纳不同合理立场，否则容易引发信任危机与合规风险。与此同时，外部制度环境会通过政策法规形成硬性约束，要求企业人工智能价值对齐规范化（Ebers，2025）。此外，Yang等（2025）指出，人工智能伦理教育与数字素养的提升，也会增强公众监督与行业问责压力，进而推动企业加大对人工智能价值对齐的投入。由此可见，社会价值的多样性、政策法律的强制性、公众意识的不断提高向企业人工智能价值对齐提出了更高的要求，企业的人工智能系统部署也必然受到社会环境因素的影响。

2. 技术因素。由于不同类型的模型在算法架构、目标设定上存在明显差异，其运行逻辑会直接影响人工智能系统与企业价值诉求之间的对齐程度。研究表明，机器的深度学习依赖于大规模数据的训练，若数据来源本身存在不平衡或系统偏见，可能直接导致企业部署的人工智能系统存在价值偏离（Shah和Sureja，2025）。不仅如此，即使是较少依赖传统训练数据的强化学习方法，也可能存在价值目标冲突的问题（Leike等，2018；Gabriel，2020；Marklund等，2025）。此外，随着人工智能系统在高风险场景中的应用，企业需要借助风险评估、算法审计和问责机制，持续识别系统可能产生的偏差，以增强人工智能治理的公平性、透明性和可追责性（Cath，2018）。总体来看，目前技术因素的研究视角仍局限于模型本身的底层算法逻辑，未来仍需进一步探索企业部署人工智能系统的技术能力、技术资源等因素对人工智能价值对齐的影响。

综上所述，现有文献研究从组织内外部因素分别讨论了对企业人工智能价值对齐的影响，但目前研究仍不全面，也尚未形成统一的分析框架，并且对不同因素之间的交互作用缺乏深入探讨。因此，为了完整刻画企业人工智能价值对齐的影响过程，有必要进一步探讨其影响效应和作用机制。

六、企业人工智能价值对齐的影响效应与作用机制

（一）企业人工智能价值对齐的影响效应

文献研究表明，既有研究探讨的影响效应主要可划分为个体、组织和社会三个层面。

1. 个体结果

（1）对个体认知的影响。企业人工智能价值对齐对个体认知的影响主要包括认知信任（Feldkamp等，2024）、认知评价（Yu和Li，2022；孔祥维等，2022）等。同时，人工智能系统在企业中的应用，尤其是在透明性、可解释性和价值约束不足的情况下，可能加剧员工对岗位替代和工作连续性的担忧，形成工作不安全感（Chung等，2025）。研究表明，当系统表现出高度的价值对齐，用户会显著增强对算法公正性的积极评价，进而提升对系统的信任（Feldkamp等，2024）。Yu和Li（2022）基于压力认知评价理论指出，企业人工智能价值对齐能引导员工将人工智能系统视为一种可控的挑战而非无法预测的潜在威胁。孔祥维等（2022）进一步指出，这种认知上的转变主要依赖于组织是否建立清晰的人工智能问责机制。

（2）对个体情感的影响。企业人工智能价值对齐对个体情感的影响主要集中在情绪耗竭（周恋等，2022）、焦虑感（Yu等，2023）等负面情绪。周恋等（2022）的研究表明，当企业人工智能系统展现出价值偏离时，不公平感这种持续的负面感受会引发情绪耗竭。与之相反，Yu等（2023）指出，具备高透明度的企业人工智能系统架构可以作为一种情感调节工具，通过降低不确定性来缓解员工对未知算法的焦虑体验。

（3）对个体态度的影响。企业人工智能价值对齐对个体态度的影响主要包括算法接受度（王戈和张哲君，2023）、持续使用意愿（Rane等，2024）、工作满意度（Tang等，2023）等。王戈和张哲君（2023）研究发现，价值对齐程度高的系统能显著提高个体对算法决策结果的接受态度。在此基础上，Rane等（2024）研究指出，这种积极的接受态度会转化为长期稳定的持续使用意愿。Tang等（2023）进一步研究发现，当员工倾向于接受并使用企业的人工智能系统时，这种嵌入工作流程的技术工具会进一步提高员工的工作满意度。

（4）对个体行为的影响。企业人工智能价值对齐对个体行为的影响主要包括创新行为（Tang等，2023）、防御性行为（Yuan等，2025）、反生产行为（Meng等，2025）、知识隐藏行为（Liu等，2025）等。Tang等（2023）实证研究发现，有效的人机信任会促进员工主动探索技术构思，展现出显著的创新行为。反之，当感知到企业人工智能系统价值不对齐或产生隐私侵犯时，员工会采取报复性或防御性行动。Meng等（2025）研究表明，价值失衡可能引发反生产行为，如针对算法的隐蔽反抗、网络闲逛或刻意磨洋工。此外，最近研究关注到了知识领域的消极行为。Liu等（2025）指出，企业人工智能系统存在的隐私风险会导致员工的防御性收缩，即为了预防个人资源外泄而刻意隐藏关键知识或提供虚假信息。基于此，Mkhize和Lourens（2025）进一步发现，企业人工智能价值对齐能有效抑制知识隐藏行为，提升个体的知识获取能力。

总体来看，现有研究主要围绕企业人工智能价值对齐对个体的认知、情感、态度与行为结果展开讨论，但对其如何影响个体动机的研究相对不足。同时，关于情感结果的研究多集中于焦虑、情绪耗竭等负面体验，缺乏对自豪、感恩等积极情感的关注。

2. 组织结果

现有关于企业人工智能价值对齐对组织的影响研究，主要围绕提升组织绩效、提高组织治理能力、改善组织社会形象等几个方面。Xia等（2026）基于证券公司样本的准自然实验表明，企业推进人工智能价值对齐实践与企业的市场估值存在显著的正向关系。Papagiannidis等（2023）通过多案例研究指出，对齐实践不仅帮助企业减少人工智能应用的负面影响，还有助于提升开

发流程的稳健性。Hiremath等(2025)研究发现,提高人工智能系统的公平性与决策过程的透明度,能够在一定程度上改进组织决策质量。Janssen等(2020)进一步强调,价值对齐是实现人工智能系统可信、可控的重要基础,缺乏有效对齐的组织更容易在算法偏差、隐私保护及责任界定等方面面临系统性风险。此外,企业人工智能价值对齐的治理实践,作为一种信号机制有利于提升利益相关者对组织的信任和支持。例如,Park和Yoon(2025)发现,企业人工智能算法的透明度可缓解公众对人工智能的负面态度,提高公众对企业的信任与支持。同时,明确人工智能系统的代理权边界,建立相对完善的审计监督机制,有助于企业降低算法滥用与违规风险(Lu,2019)。尽管已有研究从多方说明企业人工智能价值对齐实践对组织产生的积极影响,但也有研究提出,过早或过度的对齐信息披露可能侵蚀利益相关者信任(Schilke和Reimann,2025)。由此可见,企业人工智能价值对齐对组织层面的影响研究结论并不一致,其有效发挥作用的边界条件仍有待进一步探讨。

3. 社会结果

企业人工智能价值对齐的影响不限于个体及组织内部,更具有广泛的社会外部性。由于行业属性、应用场景及受众群体的差异,其影响在社会范围、表现形式上也并不相同。已有研究表明,企业招聘环节中的算法偏差可能造成系统性不利后果,损害弱势群体的平等就业机会与公正对待(王戈和张哲君,2023;Ochmann等,2024)。Capasso等(2024)进一步指出,算法化人力资源管理若缺少对隐私、非歧视与可申诉性的制度化约束,将对劳动者基本权利(如平等、隐私与工作权)构成实质性风险。而在某些关键领域,企业人工智能价值对齐会直接影响公众生命健康和社会公共安全。例如,Saeidnia等(2026)指出,在医疗情境中,当企业缺乏价值对齐与责任约束时,生成式人工智能可能被用于制造或传播医学错误信息,从而对公共健康造成现实风险。这意味着企业由于人工智能系统价值偏差所产生的错误内容会通过内容生成、传播与平台治理链条,进一步影响社会信息环境,危及公共安全底线(Park和Nan,2026)。总体来看,企业人工智能价值对齐对社会结果的影响研究视角分散,尚缺乏统一的分析框架进行整合,未来研究可以进一步围绕社会外部性的影响过程展开讨论。

(二)企业人工智能价值对齐的作用机制

1. 中介机制

(1)认知视角。从认知视角来看,既有研究的理论基础主要包括认知评价理论和社会信息加工理论。根据认知评价理论,企业人工智能价值对齐通过改变员工对技术的认知评价过程发挥作用,包括积极认知评价(Feldkamp等,2024)、消极认知评价(Chung等,2025)、挑战性和阻碍性压力认知评价(Yu和Li,2022)等。社会信息加工理论则强调个体通过对人工智能系统传递出的特定信息进行加工和解读,形成相应的情感表现(周恋等,2022;Yu等,2023)。总体来看,这一理论视角的有效性高度依赖组织信息的透明度,若信息披露不充分,个体很难从复杂的算法黑箱中获取有效信息,并进行加工判断。

(2)资源视角。从资源视角来看,企业人工智能价值对齐可以被视为组织向员工提供的非物质性资源投入或伦理承诺。现有研究表明,当人工智能系统体现隐私保护、公平和安全等价值原则时,员工更容易将其感知为组织关怀,并形成较高的组织信任和积极行为回应(Gkinko和Elbanna,2023;Tang等,2023)。相反,当人工智能系统出现价值不对齐时,员工可能因资源受损而采取报复性或防御性行为,如算法反抗、网络闲逛、磨洋工、知识共享减少和知识隐藏(Meng等,2025;Liu等,2025)。

2. 调节机制

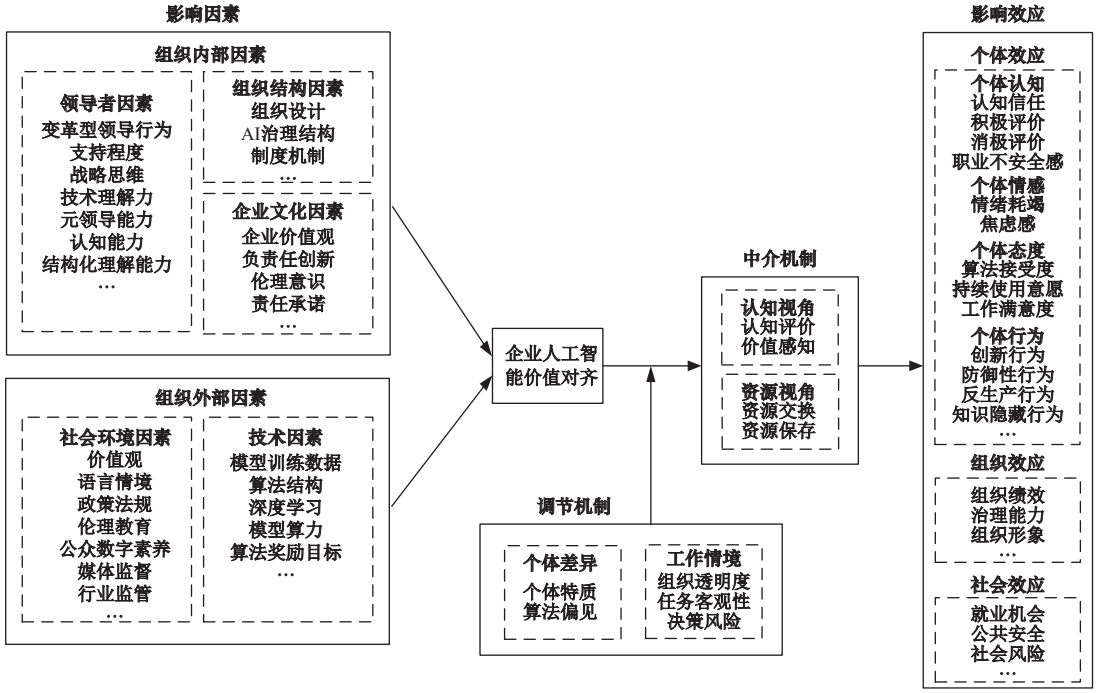
回顾已有研究,有关企业人工智能价值对齐调节机制的研究相对匮乏,少数研究引入了调

节变量，探讨企业人工智能价值对齐作用机制的边界条件，主要包括个体差异变量和工作情境变量。

（1）个体差异变量。在企业人工智能价值对齐影响效应的研究中，个体差异变量主要包括个体特质和算法偏见。研究表明，将错误视为学习机会的个体更容易在错误反馈后保持适应性反应，表现出更高的元认知控制和努力投入（Tulis和Dresel，2025）。Feldkamp等（2024）指出，当个体持有算法厌恶倾向时，即使人工智能系统表现出较高的价值对齐，其对心理安全感的提升作用也会因个体的预设偏见而打折扣。此外，在算法偏见情境中，个体的数字自我效能、技术知识和平等信念等差异，也会影响其对人工智能系统的信任、负面情绪和怀疑程度（Kim，2024）。

（2）工作情境变量。工作情境变量也是企业人工智能价值对齐发挥作用的重要边界条件。现有研究表明，组织透明度会影响员工对算法传递信息的感知和解释（Yu等，2023）。同时，任务决策场景和结果（Meng等，2025）、任务客观性（王戈和张哲君，2023）、决策类型和数据来源（Majrashi，2025）等都会影响企业人工智能价值对齐的影响效应。

总体来说，与探讨前因和结果变量相比，目前企业人工智能价值对齐作用机制的研究相对匮乏，相关研究工作仍有待进一步开展。通过对人工智能价值对齐相关文献的系统性回顾，本研究构建了企业人工智能价值对齐的影响因素、结果变量以及中介机制和调节机制的整合研究框架，如图2所示。



资料来源：根据相关文献整理。

图 2　企业人工智能价值对齐的整合研究框架

七、结论与展望

通过对大量文献进行回顾与总结，本研究发现虽然现有文献对企业人工智能价值对齐的概念、结构、前因及后果等展开了初步探索，但是仍存在一定的局限性。基于此，本文对既有研究的局限性进行总结，并对未来可能的研究方向进行展望。

（一）概念内涵的完善及本土化发展

尽管人工智能价值对齐成为人工智能伦理研究的重要议题，但目前学术界对人工智能价值对齐的定义尚未达成共识，并且部分研究者将“value alignment”译为“价值观对齐”（晋少雄和刘超，2025；吴静，2025；易显飞和高津宇，2026），缺乏统一的概念表达，不利于后续本土化发展。虽然本研究基于文献回顾与梳理对人工智能价值对齐概念进行了界定，但该定义能否充分反映其内涵并与相关概念明确区分开来，仍需研究验证。同时，现有研究在人工智能价值对齐的结构上划分不明确，虽然本研究整合透明度、正义与公平、无害、责任与隐私等结构维度，在一定程度上揭示人工智能价值对齐的本质内容，但上述结构要素仍难以充分覆盖人类社会的多元价值体系。以往研究表明，价值对齐的核心“价值”具有较强的社会属性，不同文化背景下的价值内涵并不完全一致（Gabriel和Ghazavi，2022；Jaung，2026）。现有价值对齐的结构框架多基于西方价值体系，强调个体权利、公平与正义等准则。相比之下，中国传统文化可能更看重集体利益、和谐关系与伦理秩序等社会价值。未来研究可以适当结合中国哲学，从当代管理实践出发，重新定义企业人工智能价值对齐概念的核心维度。例如，深入挖掘中国传统文化中“仁、义、礼、智、信”等伦理思想，以及当代企业所提倡的诚信、互惠等价值原则，加强概念的本土解释力。

（二）聚焦主体研究，构建多方协同的价值对齐实现机制

既有研究分别从对齐的内容、方向和过程对人工智能价值对齐的实现类型展开探索，主要聚焦于算法技术的优化和制度规范的约束（Erdélyi和Goldsmith，2018；Mökander和Floridi，2021）。然而，这些研究多以通用人工智能系统为研究对象，较少区分企业、政府、第三方审查机构、技术开发者和用户等不同行为主体在价值对齐中的责任边界与协作方式。就企业人工智能价值对齐而言，价值对齐还涉及不同职能部门之间的协同治理。尽管Carter（2020）和Shen等（2024）研究提到多方协同对于促进价值对齐的重要性，但已有研究仅提出基本设想，并未深入探讨各主体之间的协作机制如何实现。因此，未来研究可以结合协同治理理论，通过共同参与、规则协商、信息共享和责任分担形成多主体治理的合作机制，充分调动企业各层级、跨部门共同参与，构建多方协同的企业人工智能价值对齐实现机制。

（三）深化企业人工智能价值对齐的影响因素研究

目前，关于企业人工智能价值对齐影响因素的研究仍有进一步拓展的空间。现有研究多关注领导者行为和能力，较少考察领导者道德认同、风险偏好和技术焦虑等个体差异对价值对齐资源投入和决策的影响。同时，既有研究更重视外部算法演进（Shah和Sureja，2025），对企业内部数据资源、人工智能人才储备和算法审计能力等组织资源关注不足。此外，基于社会技术系统理论讨论技术—组织匹配是否对企业人工智能价值对齐产生差异化影响也是未来可供发展的方向。

（四）丰富企业人工智能价值对齐的影响效应和作用机制

1. 探索企业人工智能价值对齐的团队效应

关于企业人工智能价值对齐的影响研究，目前主要集中在个体、组织及社会三个层面，对团队的影响在现有研究中尚未得到充分关注。研究表明，企业人工智能价值对齐能显著提升个体的认知评价与算法信任（Feldkamp等，2024；Yu和Li，2022）。根据社会认知理论，这种个体的信任感往往会通过团队内部的互动转化为集体层面的认知。同时，已有实证研究发现，当企业人工智能价值不对齐时，个体会产生知识隐藏行为（Liu等，2025）或反生产行为（Meng等，2025）。在团队情境下，个体的知识隐藏不仅涉及资源流失，更会通过人际交往而影响团队内部的知识流动，损害团队的集体创新能力（Tang等，2023）。未来研究可以进一步考察，针对算法

价值偏离而产生的个体防御性行为，是否会在团队内部产生扩散效应，导致团队成员形成共同的消极认知或抵制行为。

2. 拓展企业人工智能价值对齐的作用机制

既有文献研究企业人工智能价值对齐的影响效应聚焦于个体层面的结果，导致其中介机制多围绕微观个体，探讨个体内部的认知评价作用和个体与组织的资源交换，较少关注与其他利益相关者的互动过程。依据信号理论，由于人工智能技术具有高度的复杂性与不透明性，企业与外部利益相关者之间存在着严重的信息不对称。基于此，企业的人工智能价值对齐实践也是一种向外界传递其治理能力与社会责任感的高质量信号。未来的研究可以深入分析这种信号传递如何转化为企业的信用资本，帮助企业在获取政策支持、吸引ESG投资以及提升品牌溢价等方面获得战略优势（Young等，2019；Lu，2019）。在调节机制方面，现有研究聚焦于个体差异和工作情境两类变量，未来可以进一步关注领导者、团队等多方面的影响。

（五）企业人工智能价值对齐的跨文化研究

研究指出企业人工智能价值对齐具有情境适应性，涉及不同国家和地区、不同行业的标准与价值规范。因此，在人工智能技术全球发展的进程中，跨文化背景下的人工智能价值对齐研究尤为重要。对于跨国企业而言，人工智能系统不仅要与总部的战略价值对齐，还需要包容不同地区多元化的价值立场，否则极易引发信任危机与合规风险（Habbal等，2024）。例如，在涉及隐私保护、数据公开等个人权利问题时，不同国家的法律框架与社会习俗可能产生较大的价值冲突。未来应尝试开展跨文化的比较研究，通过对比不同文化背景下企业人工智能价值对齐实践的差异，探讨如何构建具有“文化敏感性”的价值对齐系统。

主要参考文献

[1]付业辉, 僧建芬, 乐凯迪, 等. 人工智能应用与企业价值链升级: 效应、机制与情境异质性——大语言模型的文本分析证据[J]. 科技进步与对策, 2026, 43(8): 26-36.

[2]古天龙, 马露, 李龙, 等. 符合伦理的人工智能应用的价值敏感设计: 现状与展望[J]. 智能系统学报, 2022, 17(1): 2-15.

[3]晋少雄, 刘超. 由“心”及“智”: 心理学研究促进AI价值观对齐的路径探讨[J]. 心理科学, 2025, 48(4): 782-791.

[4]孔祥维, 王子明, 王明征, 等. 人工智能使能系统的可信决策: 进展与挑战[J]. 管理工程学报, 2022, 36(6): 1-14.

[5]李海舰, 赵丽. 数据价值理论研究[J]. 财贸经济, 2023, 44(6): 5-20.

[6]李思雯. 人工智能价值对齐的路径探析[J]. 伦理学研究, 2024, (5): 99-108.

[7]凌斌, 贺筱颖, 王晓辰. AI算法视域下决策建议来源对验证性信息偏差的影响[J]. 心理科学, 2025, 48(5): 1185-1196.

[8]裴嘉良, 刘善仕, 钟楚燕, 等. AI算法决策能提高员工的程序公平感知吗?[J]. 外国经济与管理, 2021, 43(11): 41-55.

[9]宋保林. 人工智能大模型价值对齐的伦理建构[J]. 伦理学研究, 2025, (3): 94-99.

[10]孙锐, 袁圆, 朱秋华, 等. 感知算法控制的双刃剑效应对零工工作者情绪耗竭的影响: 基于合法性判断视角[J]. 系统管理学报, 2024, 33(5): 1373-1385,1396.

[11]唐林垚. 公司法如何促进模型可信与价值对齐[J]. 东方法学, 2024, (2): 76-87.

[12]王戈, 张哲君. 任务客观性、情感相似度何以影响算法决策感知公平与接受度?——基于调查实验的实证分析[J]. 公共管理与政策评论, 2023, 12(6): 77-95.

[13]吴静. 人工智能价值观的动态适应对齐研究——从意图—价值—情境互构范式到智能正义[J]. 华中科技大学学报(社会科学版), 2025, 39(6): 1-10.

[14]席丹. 寻求价值对齐之路: 人工智能面临的课题与挑战[J]. 传媒, 2024, (11): 41-43.

[15]夏永红. 人工智能伦理治理范式: 从价值对齐到价值共生[J]. 自然辩证法通讯, 2025, 47(1): 1-8.

[16]闫坤如. 人工智能体价值对齐的分布式路径探赜[J]. 上海师范大学学报(哲学社会科学版), 2024, 53(4): 131-139.

[17]姚加权, 张锟澎, 郭李鹏, 等. 人工智能如何提升企业生产效率?——基于劳动力技能结构调整的视角[J]. 管理世界, 2024, 40(2): 101-116,133.

[18]矣晓沅, 谢幸. 大模型道德价值观对齐问题剖析[J]. 计算机研究与发展, 2023, 60(9): 1926-1945.

[19]易显飞, 高津宇. 幻觉的蒸馏: 生成式人工智能价值观三条对齐路径的价值风险解析[J]. 东岳论丛, 2026, 47(1): 26-34,191.
[20]曾雯, 侯凌风, 周旅军. 价值对齐: 人工智能时代的技术伦理与文明对话[J]. 贵州民族大学学报(哲学社会科学版), 2025, (6): 172-191.
[21]赵艺. “人工智能+”时代的价值对齐: 困境与治理[J/OL]. 科学学研究, 2026: 1-14.
[22]赵一骏, 许丽颖, 喻丰, 等. 感知不透明性增加职场中的算法厌恶[J]. 心理学报, 2024, 56(4): 497-514.
[23]周恋, 雷雪, 后锐, 等. 在线用工平台算法管理的消极影响和控制策略研究: 算法技术属性视角[J]. 中国人力资源开发, 2022, 39(6): 8-22.
[24]Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[A]. Proceedings of 2016 ACM SIGSAC conference on computer and communications security[C]. Vienna: ACM, 2016.
[25]Abositta A, Adedokun M W, Berberoğlu A. Influence of artificial intelligence on engineering management decision-making with mediating role of transformational leadership[J]. Systems, 2024, 12(12): 570.
[26]Acikgoz Y, Davison K H, Compagnone M, et al. Justice perceptions of artificial intelligence in selection[J]. International Journal of Selection and Assessment, 2020, 28(4): 399-416.
[27]Ali B A. AI integrated approach to achieve transformational strategic leadership and its reflections on employee engagement[A]. The International Conference on Artificial Intelligence Management and Trends[C], 2025, 87.
[28]Baker J, Cao Q, Jones D, et al. Dynamic strategic alignment competency: A theoretical framework and an operationalization[R]. Working Paper, 2009.
[29]Barabadi E, Fotuhabadi Z, Arghavan A, et al. Comparing AI and human moral reasoning: Context-sensitive patterns beyond utilitarian bias[J]. Frontiers in Artificial Intelligence, 2026, 8: 1710410.
[30]Beheshti A, Kerridge I. Understanding the artificial intelligence revolution and its ethical implications[J]. Journal of Bioethical Inquiry, 2025, 22(3): 497-505.
[31]Boncella R. AI and management: Navigating the alignment problem for ethical and effective decision-making[J]. Issues in Information Systems, 2024, 25(4): 194-204.
[32]Bostrom N. Superintelligence: Paths, dangers, strategies[M]. Oxford: Oxford University Press, 2014.
[33]Braun M, Greve M, Gnewuch U.The new dream team? A review of human-AI collaboration research from a human teamwork perspective[A]. Proceedings of the International Conference on Information Systems[C]. Hyderabad: Association for Information Systems, 2023.
[34]Brown C V. Examining the emergence of hybrid IS governance solutions: Evidence from a single case site[J]. Information Systems Research, 1997, 8(1): 69-94.
[35]Brown O, Davison R M, Decker S, et al. Theory-driven perspectives on generative artificial intelligence in business and management[J]. British Journal of Management, 2024, 35(1): 3-23.
[36]Cai Y N, Wang C. Effect of transparency of algorithmic performance management on proactive work behavior: Role of motivation to improve performance and psychological ownership[J]. International Journal of Asian Business and Information Management (IJABIM), 2024, 15(1): 1-17.
[37]Capasso M, Arora P, Sharma D, et al. On the right to work in the age of artificial intelligence: Ethical safeguards in algorithmic human resource management[J]. Business and Human Rights Journal, 2024, 9(3): 346-360.
[38]Carter D. Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role?[J]. Business Information Review, 2020, 37(2): 60-68.
[39]Cath C. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges[J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2018, 376(2133): 20180080.
[40]Chan Y E, Reich B H. IT alignment: What have we learned?[J]. Journal of Information Technology, 2007, 22(4): 297-315.
[41]Cheong B C. Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making[J]. Frontiers in Human Dynamics, 2024, 6: 1421273.
[42]Chung Y W,Im S,Kim J E,et al. Artificial intelligence awareness, career resilience, job insecurity and behavioural outcomes[J]. Australian Journal of Psychology, 2025, 77(1): 2559910.
[43]Csaszar F A, Ketkar H, Kim H. Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and

investors[J]. Strategy Science, 2024, 9(4): 322-345.
[44]de Laat P B. Companies committed to responsible AI: From principles towards implementation and regulation?[J]. Philosophy & Technology, 2021, 34(4): 1135-1193.
[45]Decker M C, Wegner L, Leicht-Scholten C. Procedural fairness in algorithmic decision-making: The role of public engagement[J]. Ethics and Information Technology, 2025, 27(1): 1.
[46]Ebers M. Truly risk-based regulation of artificial intelligence: How to implement the EU's AI Act[J]. European Journal of Risk Regulation, 2025, 16(2): 684-703.
[47]Enholm I M, Papagiannidis E, Mikalef P, et al. Artificial intelligence and business value: A literature review[J]. Information Systems Frontiers, 2022, 24(5): 1709-1734.
[48]Erdélyi O J, Goldsmith J. Regulating artificial intelligence: Proposal for a global solution[A]. Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society[C]. New Orleans: ACM, 2018.
[49]Fabris A, Baranowska N, Dennis M J, et al. Fairness and bias in algorithmic hiring: A multidisciplinary survey[J]. ACM Transactions on Intelligent Systems and Technology, 2025, 16(1): 16.
[50]Feldkamp T, Langer M, Wies L, et al. Justice, trust, and moral judgements when personnel selection is supported by algorithms[J]. European Journal of Work and Organizational Psychology, 2024, 33(2): 130-145.
[51]Ferrara E. Addressing racial bias in AI: Towards a more equitable future[A]. Czarnowski I, Howlett R J, Jain L C. Intelligent decision technologies[M]. Singapore: Springer, 2025.
[52]Firt E. Addressing corrigibility in near-future AI systems[J]. AI and Ethics, 2025, 5(2): 1481-1490.
[53]Floridi L, Cowls J. A unified framework of five principles for AI in society[A]. Carta S. Machine learning and the city: Applications in architecture and urban design[M]. Hoboken: John Wiley & Sons Ltd, 2022.
[54]Floridi L, Cowls J, King T C, et al. How to design AI for social good: Seven essential factors[A]. Floridi L. Ethics, governance, and policies in artificial intelligence[M]. Cham: Springer, 2021.
[55]Gabriel I. Artificial intelligence, values, and alignment[J]. Minds and Machines, 2020, 30(3): 411-437.
[56]Gabriel I, Ghazavi V. The challenge of value alignment: From fairer algorithms to AI safety[A]. Véliz C. The Oxford handbook of digital ethics[M]. Oxford: Oxford University Press, 2022.
[57]García-Madurga M Á, Gil-Lacruz A I, Saz-Gil I, et al. The role of artificial intelligence in improving workplace well-being: A systematic review[J]. Businesses, 2024, 4(3): 389-410.
[58]Gehman S, Gururangan S, Sap M, et al. RealToxicityPrompts: Evaluating neural toxic degeneration in language models[A]. Findings of the Association for Computational Linguistics: EMNLP 2020[C]. 2020: 3356-3369.
[59]Gkinko L, Elbanna A. Designing trust: The formation of employees' trust in conversational AI in the digital workplace[J]. Journal of Business Research, 2023, 158: 113707.
[60]Glikson E, Woolley A W. Human trust in artificial intelligence: Review of empirical research[J]. Academy of Management Annals, 2020, 14(2): 627-660.
[61]Habbal A, Ali M K, Abuzaraida M A. Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions[J]. Expert Systems with Applications, 2024, 240: 122442.
[62]Hadfield-Menell D, Dragan A, Abbeel P, et al. Cooperative inverse reinforcement learning[A]. Proceedings of the 30th international conference on neural information processing systems[C]. Barcelona: Curran Associates Inc., 2016.
[63]Hickman E, Petrin M. Trustworthy AI and corporate governance: The EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective[J]. European Business Organization Law Review, 2021, 22(4): 593-625.
[64]Hiremath S, P R, Konek S S, et al. Artificial intelligence (AI) governance in organizational decision-making: Balancing autonomy, accountability and transparency[J]. Journal of Entrepreneurship and Public Policy, 2025: 1-24.
[65]Hristova T, Magee L, Soldatic K. The problem of alignment[J]. AI & Society, 2025, 40(3): 1439-1453.
[66]Janssen M, Brous P, Estevez E, et al. Data governance: Organizing data for trustworthy artificial intelligence[J]. Government Information Quarterly, 2020, 37(3): 101493.
[67]Jaung W. Does AI value the environment? Evaluation of AI value alignment[J]. Technological Forecasting and Social Change, 2026, 225: 124550.
[68]Ji J M, Qiu T Y, Chen B Y, et al. AI alignment: A comprehensive survey[Z]. arXiv: 2310.19852, 2023.

[69]Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389-399.
[70]Josifović S. Legal and administrative frameworks as foundations for AI alignment with human volition[J]. AI and Ethics, 2025, 5(3): 3057-3067.
[71]Kim S. Perceptions of discriminatory decisions of artificial intelligence: Unpacking the role of individual characteristics[J]. International Journal of Human-Computer Studies, 2025, 194: 103387.
[72]Kumar S, Datta S, Singh V, et al. Applications, challenges, and future directions of human-in-the-loop learning[J]. IEEE Access, 2024, 12: 75735-75760.
[73]Leike J, Krueger D, Everitt T, et al. Scalable agent alignment via reward modeling: A research direction[Z]. arXiv: 1811.07871, 2018.
[74]Leslie D. Understanding artificial intelligence ethics and safety[Z]. arXiv: 1906.05684, 2019.
[75]Liu Z, Lin Q X, Tu S M, et al. When robot knocks, knowledge locks: How and when does AI awareness affect employee knowledge hiding?[J]. Frontiers in Psychology, 2025, 16: 1627999.
[76]Lu Y. Artificial intelligence: A survey on evolution, models, applications and future trends[J]. Journal of Management Analytics, 2019, 6(1): 1-29.
[77]Madanchian M, Taherdoost H. Ethical theories, governance models, and strategic frameworks for responsible AI adoption and organizational success[J]. Frontiers in Artificial Intelligence, 2025, 8: 1619029.
[78]Majrashi K. Employees' perceptions of the fairness of AI-based performance prediction features[J]. Cogent Business & Management, 2025, 12(1): 2456111.
[79]Marklund H, Infanger A, Van Roy B. Misalignment from treating means as ends[Z]. arXiv: 2507.10995, 2025.
[80]Martin K. Ethical implications and accountability of algorithms[J]. Journal of Business Ethics, 2019, 160(4): 835-850.
[81]Matthews M J, Su R K, Yonish L, et al. A review of artificial intelligence, algorithms, and robots through the lens of stakeholder theory[J]. Journal of Management, 2025, 51(6): 2627-2676.
[82]Meng Q, Wu T J, Duan W, et al. Effects of employee–artificial intelligence (AI) collaboration on counterproductive work behaviors (CWBs): Leader emotional support as a moderator[J]. Behavioral Sciences, 2025, 15(5): 696.
[83]Mkhize S, Lourens M. Elevating knowledge sharing and communication through artificial intelligence: An organizational perspective[J]. International Journal of Research in Business and Social Science, 2025, 14(4): 103-114.
[84]Mohammadi H, Bagheri A. Exploring cultural variations in moral judgments with large language models[Z]. arXiv: 2506.12433, 2025.
[85]Mökander J, Floridi L. Ethics-based auditing to develop trustworthy AI[J]. Minds and Machines, 2021, 31(2): 323-327.
[86]Norhashim H, Hahn J. Measuring human-AI value alignment in large language models[A]. Proceedings of the 7th AAAI/ACM conference on AI, ethics, and society[C]. San Jose: AAAI, 2024.
[87]Nouis S C, Uren V, Jariwala S. Evaluating accountability, transparency, and bias in AI-assisted healthcare decision- making: A qualitative study of healthcare professionals' perspectives in the UK[J]. BMC Medical Ethics, 2025, 26(1): 89.
[88]Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: What it is and how it works[J]. AI & Society, 2024, 39(4): 1871-1882.
[89]Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations[J]. Science, 2019, 366(6464): 447-453.
[90]Ochmann J, Michels L, Tiefenbeck V, et al. Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting[J]. Information Systems Journal, 2024, 34(2): 384-414.
[91]Omohundro S M. The basic AI drives[A]. Proceedings of the 1st AGI conference[C]. Memphis: IOS Press, 2008.
[92]Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[A]. Proceedings of the 36th international conference on neural information processing systems[C]. New Orleans: Curran Associates Inc., 2022.
[93]Papagiannidis E, Enholm I M, Dremel C, et al. Toward AI governance: Identifying best practices and potential barriers and outcomes[J]. Information Systems Frontiers, 2023, 25(1): 123-141.
[94]Park H J. Patient perspectives on informed consent for medical AI: A web-based experiment[J]. Digital Health, 2024, 10.
[95]Park K, Yoon H Y. AI algorithm transparency, pipelines for trust not prisms: Mitigating general negative attitudes and

enhancing trust toward AI[J]. Humanities and Social Sciences Communications, 2025, 12(1): 1160.
[96]Park S, Nan X L. Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation[J]. AI & Society, 2026, 41(2): 1501-1515.
[97]Pelletier C, Raymond L. Investigating the strategic IT alignment process with a dynamic capabilities view: A multiple case study[J]. Information & Management, 2024, 61(4): 103819.
[98]Raji I D, Smart A, White R N, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing[A]. Proceedings of 2020 conference on fairness, accountability, and transparency[C]. Barcelona: ACM, 2020.
[99]Rane N, Choudhary S P, Rane J. Acceptance of artificial intelligence: Key factors, challenges, and implementation strategies[J]. Journal of Applied Artificial Intelligence, 2024, 5(2): 50-70.
[100]Russell S, Dewey D, Tegmark M. Research priorities for robust and beneficial artificial intelligence[J]. AI Magazine, 2015, 36(4): 105-114.
[101]Russell S J. Human compatible: Artificial intelligence and the problem of control[M]. London: Allen Lane, 2019.
[102]Sabherwal R, Hirschheim R, Goles T. The dynamics of alignment: Insights from a punctuated equilibrium model[J]. Organization Science, 2001, 12(2): 179-197.
[103]Saeidnia H R, Jahani S, Ghiasi N, et al. Generative AI and health misinformation: Production, propagation, and mitigation—a systematic review[J]. BMC Public Health, 2026, 26(1): 693.
[104]Salaheldin S, Hussein S. The determinants of AI adoption and its impact on employee engagement: Evidence from Egyptian organizations[J]. International Journal of Management and Applied Research, 2025, 12(2): 45-67.
[105]Sarkar A, Faik I. Structural transparency of societal AI alignment through institutional logics[Z]. arXiv: 2602.08246, 2026.
[106]Schilke O, Reimann M. The transparency dilemma: How AI disclosure erodes trust[J]. Organizational Behavior and Human Decision Processes, 2025, 188: 104405.
[107]Seghid N, Iqbal F, Al-Room K, et al. Emerging threats in AI: A detailed review of misuses and risks across modern AI technologies[J]. Frontiers in Communications and Networks, 2026, 6: 1727425.
[108]Shah M, Sureja N. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions[J]. Archives of Computational Methods in Engineering, 2025, 32(1): 255-267.
[109]Shen H, Knearem T, Ghosh R, et al. Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions[Z]. arXiv: 2406.09264v1, 2024.
[110]Sorensen T, Jiang L W, Hwang J D, et al. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties[A]. Proceedings of the 38th AAAI conference on artificial intelligence[C]. Vancouver: AAAI, 2024.
[111]Stahl B C, Antoniou J, Ryan M, et al. Organisational responses to the ethical issues of artificial intelligence[J]. AI & Society, 2022, 37(1): 23-37.
[112]Stanikzai M E, Mittal E. Strategic alignment of organizational structure based on decisions for sustainable organizational performance: A bibliometric-systematic literature review[J]. Cogent Business & Management, 2025, 12(1): 2560650.
[113]Tamkin A, Brundage M, Clark J, et al. Understanding the capabilities, limitations, and societal impact of large language models[Z]. arXiv: 2102.02503, 2021.
[114]Tang P M, Koopman J, Mai K M, et al. No person is an island: Unpacking the work and after-work consequences of interacting with artificial intelligence[J]. Journal of Applied Psychology, 2023, 108(11): 1766-1789.
[115]Tulis M, Dresel M. Effects on and consequences of responses to errors: Results from two experimental studies[J]. British Journal of Educational Psychology, 2025, 95(1): 143-161.
[116]Übellacker T. Making sense of AI limitations: How individual perceptions shape organizational readiness for AI adoption[Z]. arXiv: 2502.15870, 2025.
[117]Van de Wetering R. Dynamic enterprise architecture capabilities and organizational benefits: An empirical mediation study[Z]. arXiv: 2105.10036, 2021.
[118]Von Zahn M, Liebich L, Jussupow E, et al. Knowing (not) to know: Explainable artificial intelligence and human metacognition[J]. Information Systems Research, 2025, doi: 10.1287/isre.2024.1431.

[119]Watson E, Viana T, Zhang S J, et al. Towards an end-to-end personal fine-tuning framework for AI value alignment[J]. Electronics, 2024, 13(20): 4044.
[120]Wiener N. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers[J]. Science, 1960, 131(3410): 1355-1358.
[121]Xia H S, Chen H, Zhang J Z, et al. Exploring the impact of responsible AI governance on corporate performance: A quasi-natural experiment[J]. Technological Forecasting and Social Change, 2026, 223: 124425.
[122]Yang J F, Xie W Y, Ni J J. A framework for AI ethics literacy: Development, validation, and its role in fostering students' self-rated learning competence[J]. Scientific Reports, 2025, 15(1): 38030.
[123]Young M M, Bullock J B, Lecy J D. Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration[J]. Perspectives on Public Management and Governance, 2019, 2(4): 301-313.
[124]Yu L R, Li Y. Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort[J]. Behavioral Sciences, 2022, 12(5): 127.
[125]Yu L R, Li Y, Fan F. Employees' appraisals and trust of artificial intelligences' transparency and opacity[J]. Behavioral Sciences, 2023, 13(4): 344.
[126]Yuan Y, Shi Y, Su T, et al. Resistance or compliance? The impact of algorithmic awareness on people's attitudes toward online information browsing[J]. Frontiers in Psychology, 2025, 16: 1563592.
[127]Yudkowsky E. Artificial intelligence as a positive and negative factor in global risk[A]. Bostrom N, Ćirković M M. Global catastrophic risks[M]. Oxford: Oxford University Press, 2008.
[128]Zárate-Torres R, Rey-Sarmiento C F, Acosta-Prado J C, et al. Influence of leadership on human-artificial intelligence collaboration[J]. Behavioral Sciences, 2025, 15(7): 873.

AI Value Alignment in Enterprises: Research Progress and Future Prospects

Wang Yanfei[1], Zhang Xiaonian[1], Zhu Yu[2]
(1. *School of Business Administration*, *South China University of Technology*, *Guangzhou* 510641, *China*;
2. *School of Management*, *Jinan University*, *Guangzhou* 510632, *China*)

Abstract: The rapid development of artificial intelligence (AI) technology has brought significant opportunities for enterprise management, while also posing new governance challenges. As a critical governance pathway for mitigating value deviations, ethical risks, and security risks, AI value alignment in enterprises has garnered increasing attention. In contrast, domestic research in this field is still in its infancy. In view of this, this paper systematically reviews and synthesizes existing literature, clarifies the conceptual connotation, structural dimensions, and measurement approaches of AI value alignment, constructs a mechanism for its realization, and sorts out the antecedents and consequences of AI value alignment in enterprises to form an integrated understanding framework. Finally, future research directions in this field are prospected. This paper contributes to a deeper understanding of AI value alignment and provides a reference for empirical research and practical applications of AI value alignment in enterprises in the domestic context.

Key words: AI; value alignment; ethics of AI; corporate governance

(责任编辑:宋澄宇)