

• 专题：人工智能伦理治理的理论与实践 •

编者按：

人工智能的快速发展与广泛应用引发了深刻的伦理问题，受到社会各界的普遍关注，对人工智能进行有效治理的呼声日益高涨。从整体上看，目前关于人工智能治理的研究主要包括伦理治理与法律治理两个方向。相对于法律治理而言，伦理治理具有更强的灵活性与敏捷性，因此伦理治理在未来很长一段时期内将是人工智能治理的重要组成部分。加强人工智能伦理治理的理论研究，特别是注重从实践的角度强化学术研究的现实针对性，是实现伦理治理目标的基本前提。

本专题第一篇夏永红的论文提出了“价值共生”伦理治理范式，尝试超越价值对齐范式，并以“相互受益原则”与“相互承认原则”作为价值共生范式的核心原则，为人工智能伦理治理提供了一种新颖进路。第二篇杨进、杜严勇的论文从动机上的虚伪性与结果上的虚假性两个层面探讨了作为行动的人工智能伦理漂洗的内涵，从动机因素与外部因素分析了伦理漂洗的成因，提出了伦理漂洗的治理策略。第三篇张学义、周洲等人的论文运用价值敏感设计的理论与方法，通过调查研究测量公众对人工智能的价值敏感性的感知程度，得出了公众对不同价值的敏感程度排序，并对比了相关的影响因素，为更好地开展人工智能伦理治理工作提供了重要的参考依据。

(专题策划：杜严勇)

人工智能伦理治理范式：从价值对齐到价值共生

The Paradigm of Ethical Governance of Artificial Intelligence: From Value Alignment to Value Symbiosis

夏永红 /XIA Yonghong

(北京师范大学文理学院, 广东珠海, 519087)
(Faculty of Arts and Sciences, Beijing Normal University, Zhuhai, Guangdong, 519087)

摘要：当前人工智能伦理治理的主导范式是价值对齐，其目标是确保机器价值与人类价值一致。价值对齐范式主要采取了表征主义和行为主义的AI方案，但这些方案因为面临着常识问题的挑战，难以精准捕捉和编码复杂的人类价值观。为了解决常识问题，需要引入具身-生成AI的技术方案，让它可以把握世界中的相关性，并可以自下而上地自主生成价值观。然而，如果这种自主生成的机器价值观敌对于人类，则有可能给人类带来生存风险。有鉴于此，本文提出了一个“价值共生”的替代范式，旨在实现机器价值

基金项目：国家社会科学基金一般项目“预测心智框架下4E认知的核心论题研究”(项目编号: 21BZX044); 北京师范大学珠海校区科研启动项目“人工智能自主性问题的哲学研究”(项目编号: 28803-310432103)。

收稿日期：2024年3月2日

作者简介：夏永红(1985-)男, 云南临沧人, 北京师范大学文理学院副研究员, 研究方向为认知科学哲学、科技伦理学。
Email: yonghongx@126.com

与人类价值的和谐共生,它包含了两条AI设计原则:生存利益上的相互受益和价值观上的相互承认。

关键词: 人机对齐 生存风险 具身-生成AI 相互承认 常识问题

Abstract: The current dominant paradigm in the ethical governance of artificial intelligence is Value Alignment, which aims to ensure that machine values are consistent with human values. The Value Alignment paradigm has mainly adopted representationalist and behaviourist AI approaches, but these approaches are difficult to accurately capture and encode complex human values because of the challenge of the commonsense knowledge problem. In order to solve this problem, technological solutions based on embodied-enactive AI need to be introduced to grasp the relevance in the world and autonomously generate values from the bottom up. However, if such autonomously generated machine values are hostile to humans, it may pose an existential risk to humans. Given this, this paper proposes an alternative paradigm of Value Symbiosis, which aims to achieve a harmonious symbiosis between machine and human values. It consists of two AI design principles: mutual benefit in survival interests and mutual recognition in values.

Key Words: Human-machine alignment; Existential risk; Embodied-enactive AI; Mutual recognition; The commonsense knowledge problem

中图分类号: TP18; B82 DOI: 10.15994/j.1000-0763.2025.01.001 CSTR: 32281.14.jdn.2025.01.001

导 言

虽然人类与技术一直在协同演化,但技术常被视作仅能执行操作者意图的消极工具。然而,到了人工智能(AI)时代,技术似乎开始显现自己的目的。于是,为了让AI继续作为趁手的工具,避免其自主性带来的风险,就必须考虑:如何让它的目的与人类目的始终保持一致?这就是所谓的价值对齐问题(the Problem of Value Alignment)。价值对齐已成为当下AI伦理治理的主导范式,因为它规定了什么样的问题是合理的(即如何实现人机价值对齐),应当使用什么样的技术方法来解答问题(即表征主义和行为主义的AI方法),而这两者恰恰构成了一个范式的主要要素。^[1]

然而,价值对齐范式面临诸多挑战:价值对齐在价值上是否可取?在技术上是否可能?有鉴于此,本文提出了一个被称为价值共生(Value Symbiosis)的替代范式。这一新范式的初衷是:由于人类价值的复杂性,不应该也不可能寻求机器伦理向人类伦理的单向对齐,而应探索人类与机器的价值共生,让它们相互受益和相互承认。本文的写作将围绕下述两个目标展开:首先,在第一、二节中剖析和批判价值对齐范式;

其次,在第三节中探索价值共生范式。

一、价值对齐范式

价值对齐范式的中心问题是价值对齐问题。克里斯汀(Brian Christian)将这个问题概括为:“如何确保AI模型可以捕捉到我们的规范和价值,理解我们的意思或意图,做我们想做的事情。”^[2]显然,价值对齐的目标就是让AI模型“做我们想做的事情”,而“捕捉我们的规范和价值”和“理解我们的意思或意图”都仅仅是实现这个目标的手段。然而,一个AI模型未能“做我们想做的事情”可能源于两个相反的原因:一是在当下的有限AI语境下,由于AI不够智能,只能执行人类的字面指令,无法理解我们的真正的意思或意图;二是在未来的超级智能语境下,由于AI太过智能,并已拥有自主的目的,虽然它理解了我们的意思或意图,但却只遵守自己的意思或意图。当下对价值对齐的讨论往往将两种语境混为一谈,然而,它们引发的是两个完全不同的问题。本文将把有限AI语境下的问题称为价值对齐问题,并在第一、二节中进行处理;而把超级智能语境下的问题称之为价值共生问题,将在第三节进行处理。

价值对齐问题常被形象地称之为魔法师的

学徒问题、迈达斯国王问题或最大化曲别针问题等。维纳（Norbert Wiener）在讨论自动化的伦理后果时，曾引述歌德的叙事诗《魔法师的弟子》中的故事：魔法师的弟子施咒让一群扫帚帮他用水桶取水，却忘了让扫帚停止的咒语，导致屋内水漫金山，直至魔法师归来才解除了魔咒。^[3]另外，罗素也引述了一个古希腊神话故事：迈达斯国王许愿他所触碰之物皆变黄金，最终许愿虽然灵验，却未料连食物及亲女儿也一并化为金子。^[4]波斯特洛姆（Nick Bostrom）则设想了一个“曲别针最大化”（Paperclip Maximizer）思想实验：一个超级智能被编程令其尽可能地生产最多的回形针，最终，地球及宇宙的全部可用资源全数被转化为回形针。（[5]，p.153）这些故事都被用来警示AI的一种潜在风险：如果AI仅仅按照字面意义去理解和执行人类指令，却没有把握人类指令背后所隐含的价值观，那么一个看似无害的目标足以导致灾难性的后果。

深入分析这些故事，不难发现风险的真正根源在于AI不能完全“理解我们的意思或意图”。根据格莱斯（Paul Grice）的言语行为理论，表达的意义总是超出了它的字面意思，一个表达出来的意图也总是隐含了其他未表达的意图。^[6]对于一个正常人，取水的意图总是隐含了“不要过量取水”的意图；点物成金的意图也隐含了“不应无差别地点物成金”的意图；曲别针最大化的意图同样隐含了“不应把生存资源也制成曲别针”的意图。任何成功的行动不仅要满足字面意图，同时也应满足相关的隐含意图。上述故事中的魔法或智能能动者由于未能读取这些隐含意图导致行动失败。在人类社会中，类似的无法识别隐含意图的个体，常被认为不懂眼力见儿，而非过度智能。因此，价值对齐问题根源并非AI太过智能，而是不够智能。

价值对齐问题的解题目标是：确保AI模型在设计时就以人类的方式行事。维纳在自动化技术诞生之时就已强调：“我们最好确定植入在机器中的目的是我们真正想要的目的”。^[3]波斯特洛姆提出使用动机选择方法来设计AI，即通过设定目标系统促使AI以有益于人类的方式

行事。为此，波斯特洛姆提出了两种方法：一是直接规定（direct specification），将人类价值观编码为显性的规则并植入机器，从而可以主导AI的行动动机；二是间接规范（indirect normativity），通过让AI观察人类行为、学习人类文化和了解人类心理，由此来把握并推断出符合人类集体意愿的价值规范。（[5]，pp.263-265）另外，罗素提出了一种可证明有益的人工智能（provably beneficial AI）的方案，基于三个方针：“1. 机器的唯一目标是最大限度地实现人类的偏好。2. 机器最初不确定这些偏好是什么。3. 关于人类偏好的最终信息来源是人类行为。”（[4]，p.182）罗素认为，通过强化学习等技术手段，AI可以通过观察和互动来不断学习和更新其对人类偏好的理解，从而实现机器价值与人类价值的对齐。

上述的解题方案展示了价值对齐范式的两种技术方法：表征主义和行为主义。表征主义方法，如符号AI、深度学习等，旨在通过AI系统来表征人类的价值和伦理规则。符号AI使用逻辑框架来形式化道德决策的规则；深度学习则通过训练标注了价值判断的数据来学习人类价值观。这些方法都要将人类相关的背景常识转化为原子表征（如符号AI）或分布式表征（如深度学习），从而构建出一个人类常识数据库。行为主义的方案侧重于通过环境互动或反馈机制来学习人类价值观。目前最主要的行为主义算法就是基于人类反馈的强化学习（RLHF），它在设定一个恰当的奖励函数之后，通过人类对AI模型的输出施加反馈，奖励符合人类期望的行为，惩罚不符合的行为，从而训练AI模型学习与人类价值观一致的行为。

二、从常识问题到具身-生成AI

前述的价值对齐范式表现为：设定一种单一的机器伦理，以向单一的人类伦理的单向对齐。对此，很多论者都提出过这些担忧：首先，并不存在一种单一的人类价值观，甚至同一个时代的同一族群中也存在相互冲突的道德学说，因而难以提炼出一套普遍适用的机器伦理框架；其

次,人类价值观经常随时代演变,过去被人尊崇的价值往往在今天不合时宜,这让价值对齐的标的更加难以把握;^[7]最后,如果强行将机器伦理与人类伦理对齐,有可能会强化目前的主导价值观(往往是西方白人中产男性的价值观),从而进一步边缘化其他非主流价值观;^[8]除此之外,对单一的机器伦理的追求,也有可能导致强化操作者的权力,导向可怕的敌托邦。^[9]

撇开伦理层面的考量,价值对齐在技术上也面临常识问题的挑战。根据德雷福斯(Hubert L. Dreyfus)的概括,常识问题即“如何存储和访问人类似乎知道的所有事实”。([10], p.78)常识是普通人在日常实践中所需知道的事实的集合;如果机器要理解自然语言或做出明智行动,就需掌握大量人类常识。然而,由于人类常识大多以技能知识(know-how)而非命题知识(know-that)的形式存在,而AI恰恰难以把握技能知识,所以在模拟人类行动时就将面临困境。

德雷福斯进一步区分了两种不同的常识问题:常识理解问题和转换相关性问题。两个难题分别产生于AI的智能行动的两个步骤:首先是AI对常识的组织和把握,其次是基于常识对行动的合理规划。类似地,价值对齐也会涉及到两个相同的步骤:首先,把人类价值观嵌入到AI模型中;其次,让AI基于这些价值观合理地行动。因此,价值对齐也必然会遭遇两种常识问题。

常识理解问题即“计算机如何建造才能存储和访问大量的关于人类及其世界的信念?”([10], p.79)计算机要具有人类的常识理解,就必须具有这些日常信念。因为价值对齐的技术实现依赖于计算机对人类价值观的全面把握,所以遭遇常识理解问题将不可避免。

首先,大量人类常识都没有被记载在文本中。诸如人生经验和人情世故等人际交往规范,多为默会规则,从未被详细记录在任何伦理手册之中。比如,电梯里我与陌生人应保持何种距离?拥挤的公交车上又应如何自持?这些常识源于日常生活中的潜移默化,而非事无巨细的决疑手册。因此,仅向AI模型植入道德框架或喂入文本数据的表征主义方法,并不能捕捉

到大部分人类常识。

其次,很多人类常识的习得体现在与他者和世界的具身交互之中。比如护理中的肢体动作的细节,体现了同情、耐心等价值观,要求照护者根据被照护者的暗示灵活调整动作。这些技能的学习源于交往中的试错和反馈。然而,基于强化学习的人类反馈方式,如奖励、评注、演示、对话等,相较于涉及多感官模态、动态持续的具身交互过程,显得极为单调,因此也限制了人类向AI的价值嵌入。我们实在难以想象如何向一个缺乏身体的AI传授护理常识。

如果说常识理解问题限制了把人类价值观和常识嵌入AI,那么变化相关性问题则限制了AI基于人类价值观作出合理的行动。德雷福斯将转换相关性问题定义为:当一个能动者在世界中行动的时候,它如何才能知道什么常识是相关的,什么是不相关的,从而迅速做出推理和决策。([10], p.82) 当一个被嵌入人类价值观的机器人在世界中行动时,它也同样会遭遇变化相关性问题。

让我们考虑丹尼特(Daniel Dennett)提出的一个著名例子。假设我们对机器人1发出指令:进入一个放着定时炸弹的房间,取出一块备用电池。由于炸药和电池都在一个小车上,机器人推车取出电池的时候也附带推出炸弹。于是,炸弹爆炸了。这个例子与前面所举的几个价值对齐事例十分相似,都表明机器难以理解人类字面指令背后的隐含意图。于是,为了实现价值对齐,我们向机器人植入相关常识,开发了更先进的机器人2和机器人3。机器人2可以推演出一个行动的潜在后果,从而避免将炸弹一起带出。但它无法判断哪些后果与其目标相关,因而开始分析无关紧要的结果,如推出车子后房间颜色和车轮的变化,结果在它深陷无尽的推演时,炸弹再次爆炸。机器人3能区分哪些后果与任务目标相关,哪些不相关,但在进行这些巨量计算时,炸弹又一次爆炸了。^[11]

这个例子生动地展示了在一个复杂、动态的现实世界中,让AI完全与人类目标对齐的难度。AI需要判断现实世界中的相关性,这种判断的复杂性往往超出了其计算能力的极限。也

正是因为这个原因，目前的AI模型依然主要限于在受控的虚拟或网络世界中行动，很少能大规模地被用来规划现实世界的具身行动。

然而，虽然常识问题对于当下的有限AI构成了巨大挑战，但人类却能有效地避免于这些困扰。德雷福斯指出，这在很大程度上源于人类拥有身体。^[12]当计算机处理日常问题时，它所面临的潜在相关事实是无限的，必须判断哪些事实重要，哪些事实无聊，然后才能基于相关的事实进行规划。计算机恰恰缺乏这种判别相关事实的能力。人类之所以可以轻松判别相关事实，是因为我们在世界中有本质性的利益 (interests)，这使我们能够识别哪些操作是有意义，哪些无意义。这些利益正是基于我们的身体需要而产生的。因此，正如德雷福斯所指出的，要克服常识问题，必须构造一个人类式的身体模型，让它具有我们的“需求、欲望、快乐、痛苦、活动方式、文化背景等”。([13], pp.272-273) 这需要AI领域进行一次范式转换。

具身和生成AI (embodied and enactive AI) 代表了这种新的AI范式。其中，具身AI的核心信念是：真正的智能是从身体与环境的交互中突现的。^[14]普菲尔 (Rolf Pfeifer) 和邦加德 (Josh Bongard) 认为，具身AI通过在不同的时间尺度上来模拟身体与环境的交互来实现智能。比如，在此时此处 (here-and-now) 尺度上模拟身体与环境之间的感觉运动回路；在发育尺度上模拟个体发育中的学习过程，在演化尺度上模拟代际间的适应过程。([15], pp.82-85) 通过模拟身体与环境在不同时间尺度上的持续交互，具身AI试图让智能自下而上地突现出来，并使AI能够理解日常世界的意义。如具身AI的先驱布鲁克斯 (Rodney Brooks) 所说，只有一个具身的智能能动者，才能有效地应对现实世界，并为系统的内部运作赋予意义。^[16]

然而，具身AI仍有局限。尽管它模拟了身体与环境之间的感觉运动协调，但却未能模拟人类身体的一个本质特征：脆弱性。它对AI的意义理解极为重要。因此，弗洛伊斯 (Tom Froese) 和齐姆克 (Tom Ziemke) 提出了生成AI来补充具身AI。他们引入了生成主义的一个

重要洞见：生命系统的不稳定性是一切价值和意义的终极来源。^[17]正是因为生命是脆弱的，它必须不断地进行新陈代谢、调节周围环境的适应度，以维持自身的同一性。由此，它也需要区分有利于和不利于自身的东西，而这正是价值与意义的根本来源。弗洛伊斯认为，只有生成AI才具有世界常识，识别与自身目标相关或不相关的事实或事物，从而有效克服常识问题。^[18]

因此，要让智能系统可以内在地理解世界，并具有自身的目的和规范，就必须将其设计成为一个具身的、脆弱的系统。根据弗洛伊斯和齐姆克的观点，设计这样一种具身-生成AI要遵循两个原则：首先，系统可以在某种描述层次上产生并维持自身的同一性；其次，它可以根据生存约束积极地调节其感觉运动交互。^[17]这意味着我们不能再把智能视为一种独立于实现者的可多重实现的软件系统，而是要为其赋予一个物理身体，使其具有和有机体相似的不稳定性。这种具身-生成视野呼应了在当前AI前沿领域提出的有死计算 (Mortal Computation)，^[19]即通过创造一个依赖于物理实现的、具有自适应和自创生能力的物理计算系统，为我们创造超级智能铺平道路。

三、价值共生范式

面对价值对齐范式的内在困难，我们急需探索一种能够实现超级智能的具身-生成方案。然而，这种方案也会带来新的困境：由于超级智能的价值观是自下而上地产生的，它的价值观可能不会与人类对齐，反而会形成自主独立的目标与价值观。如果它的价值观包含危害或毁灭人类的意图，那么人类将面临极大的生存风险 (existential risk)。因此，尽管我们试图通过具身-生成AI来克服价值对齐问题，但这反而可能带来新的风险，即引发前文预告的价值共生问题：如何确保AI模型的价值观倾向于与人类共生，而不是危及人类的生存。

这一问题预示着一种AI伦理治理的新范式。在此，价值对齐问题已经被消解了，有意义的

科学问题转换为价值共生问题。因为机器已经具备自主产生的价值观,我们既不能向其植入道德代码或文本,也不能仅通过行为反馈来塑造其价值观。我们需要考虑的是:如何在一个共享的世界中与之共存。本文将论证,价值共生范式的核心是两条原则:

1. 相互受益原则:机器与人类可以因为彼此的存在而相互受益,以对方的繁荣作为自己繁荣的条件。

2. 相互承认原则:机器与人类都在价值上承认对方,不仅承认对方的工具价值,也承认对方的内禀价值。

相互受益原则试图从生物学上化解人类的生存风险。人机价值共生的基础是人机共生,即人类与机器在生存利益上的互利互惠。早在计算机诞生不久的1960年,利克莱德(Joseph Licklider)就构想过一种人机共生的前景:人类设定目标、提出假设并进行评估,而计算机将完成例行工作和数据管理任务。^[20]此后,人机共生成为了一个被广泛使用的概念,它大致可以被定义为这样一种关系:“人类和机器都可以从中相互受益,因为双方都会随着时间推移变得更加智能。如果没有人类,机器根本无法存在,它需要不断输入信息才能正常运转,而人类则在很多方面依赖于机器来提高效率,例如在计算方面。”([21], p.290)因此,人机共生就意味着人类与机器对于实现各自实用目标(Pragmatic Goals)都互有工具价值,也就是说,双方都能从对方的存在中获得生存利益。

在有限AI的语境下,人机共生相对直观,因为AI主要作为执行人类指令的工具,但在超级智能语境下,这种共生关系就难以维系。当机器拥有自主的目标和价值,它就成为一种拟生物的能动者,不再仅仅是人类目标的执行者。AI可能发展出与人类生存利益相悖的工具目标,从而给人类带来生存威胁。这种威胁不再是价值对齐问题中因为AI过于愚蠢导致的,而更像是不同生物种群间的竞争,因为超级智能开始具有自身的生态位和生存利益。

因此,在超级智能的语境下,有必要更新利克莱德的人机共生理念。新的共生理念不应

再基于人类目标主导下的人机任务分工,而应致力于实现人类与机器的目标融合。这意味着人类与机器的目标不仅不相互排斥,反而会相互推进。为此,可以考虑采取两种不同的AI设计准则:生态位并行和技能互补。

生态位并行准则的灵感来自于竞争排除原理:当两个物种占据相同的生态位的时候,它们之间的资源竞争可能导致其中一个物种的消亡。^[22]为了避免与机器的直接生存竞争,我们需要确保人类与机器占据不同的生态位。比如,让机器的栖居空间与人类不同,我们通常喜欢占据阳光灿烂的原野,但却可以让机器偏好尘沙漫天的沙漠。此外,我们也可以把生态位的概念扩展为社会生态位,要求人类与机器占据不同的社会生态位。比如,将机器分配到体力劳动或危险作业,如搬砖、扫地、打螺丝,而人类则从事创造性较强的工作,如写诗、画画、搞哲学。技能互补准则要求双方具备互补的技能,能够提供对方无法独立生产的重要资源或服务。比如,机器可以提供先进的数据处理能力,而人类则可以生产机器赖以运行的电力。这意味着我们不能让超级智能掌握所有资源的生产,也不能使其成为比人类更通用的智能体,以免它完全脱离对人类的依赖。

人机共生的第二个原则,即相互承认原则,则旨在从价值论上化解人类的生存风险。人机价值共生的理念之所以不同于简单的人机共生,在于下述考虑:在超级智能的语境下,单纯基于生存利益的人机共生并不足以化解所有生存风险,因为无论是人类还是机器的行动,很多时候可能并非以实用目标为导向,而是以价值目标为导向,比如追求审美、宗教或政治信仰。

历史上很多宗教战争和国际政治冲突,并非单纯为了物质利益的争夺,而是源于意识形态的敌对。如霍耐特(Axel Honneth)所述,人类社会中的冲突不仅仅出于争取生存资源,很多时候也为了争取对方的承认。^[23]这种为承认而斗争可能会发生在人类与机器之间,甚至在机器与机器之间。由具身-生成进路所生产的机器种群将拥有自身的周围世界(环境)和共同世界(社群),故而其在与世界交互过程中自

下而上产生的价值观将是高度情境化（situated）和在地化（local）的。因此，就像不存在同质化的人类价值观一样，也不存在一种单一的机器价值观。当这些抱有不同价值观的人类和机器种群（或族群）相遇时，价值观上的误解和蔑视将会导致冲突和斗争。

因此，当机器成为拟主体的存在后，要实现人类与机器的永久和平，仅靠物质利益上相互依赖是不够的，还需相互承认对方的独特目的、价值观和身份。如果说价值对齐范式预设了一种普遍主义的人类价值，那么价值共生范式则主张每个机器和人类社区都有其独特的、内禀的价值。这反映了两种政治哲学的对立：价值对齐范式偏向普遍主义的平等政治，要求忽视人与人之间的差异；而价值共生范式则更贴近泰勒（Charles Taylor）所辨识的差异政治，主张尊重和承认每一个个体和族群的特殊性。（[24]，p.305）鉴于未来社会将日益多元化，为了实现人与人、人与机器、机器与机器的互利共生，差异政治将是一种更可取的政治模式。

相互承认并非两个孤立个体之间的抽象认可，而是依赖于双方跨越差异的公共经验或规范，也即主体间性。胡塞尔最早引入这个概念用以表达这样的事实：我们对世界的经验不是私人的，而是公共的。^[25]而在海德格尔那里，主体间性则意味着一种与他人共在的存在方式，我们总是一起置身于一个共同世界之内。^[26]正是凭借主体间性，我们才得以走出自我的孤岛，让相互理解和相互承认得以可能。当我们承认一个人的文化身份或成就时，我们总是在某种共同的文化框架之内承认他。比如，当我们承认一个人的诗歌成就，我们也就和这个人共享了同样的社会规范——在这种规范中，诗歌创作是重要的活动，而不是毫无意义的卖弄。正是通过主体间性，承认才具有深度，而不会成为一种空洞的姿态。

因此，主体间性构成了相互承认的可能性

条件。它允许不同的个体和社群形成独特的价值观，同时又确保这些不同价值观之间的相互承认和理解。把这一理念应用于AI设计，意味着我们需要在人类与机器、机器与机器之间建造类似主体间性的东西，我们可以把它称之为（人类与机器之间的）跨主体性^[27]和（机器与机器之间的）机器间性^①。一种安全的具身-生成AI不仅要具有一个脆弱的身体，而且要让它具有同感的能力，与其他能动者共享某种共同世界，从而在陌生的价值观之间搭建起承认和理解的桥梁。在这个意义上，主体间性、跨主体性和机器间性将是未来人机和平的拱顶石。

结 语

价值对齐范式旨在创建单一的机器伦理，并向单一的人类伦理单向对齐，而价值共生范式则提倡构建多元的机器伦理，与多元的人类伦理互利共生。价值对齐既不可能——因为面临常识问题的困扰，也不可取——因为无法捕捉人类价值观的多元性和动态性；而价值共生则试图促成情境化和在地化的机器伦理与人类伦理相互承认。两种范式实际上预设了不同的形而上学：价值对齐范式延续了人本主义哲学，倡导一种普遍的人类价值观，并主张把人类的自主和尊严凌驾于其他非人类存在者之上；而价值共生范式则采纳了后人类主义哲学，认为人类价值并非既成、固定的规范，而是在不同的历史、文化和技术情境中不断生成的。

本文并不准备在两种范式之间进行裁决，因为两种范式基于不同的技术语境。我们呈现的是一个可称为“创新-风险螺旋”（Innovation-Risk Spiral）的现象：有限AI的应用引发价值对齐问题；为了解决该问题，需要发展超级智能；超级智能的应用又带来价值共生问题……尽管本文提出了两条价值共生原则来解决价值共生问题，但这个问题解决后依然可能引发新的风险，比

①赵汀阳提出了一种沟通不同文化社群的跨主体性，并把它应用到对AI能动者的讨论之中；类似地，我们可以提出一种沟通不同机器社群的机器间性。如果主体间性涉及到人类之间的沟通行动，那么跨主体性和机器间性则分别涉及到人类与机器、机器与机器之间的沟通行动。

如人类的自我退化和淘汰。在这个无尽的螺旋中,风险与创新不断交织,创新可以化解旧的风险,但却会带来新的风险。于是,我们就面临一个决断:要么采取技术保守主义,完全遏制技术创新以避免创新-风险螺旋,逃遁到一个“无风险”的避风港;要么走向技术加速主义,果断推进技术创新,接受创新-风险螺旋,在创新中调整风险。无论做出何种决断,我们都必须勇于承担后果,任何犹豫都无济于事。

[参考文献]

- [1] 托马斯·库恩. 科学革命的结构[M]. 金吾伦、胡新和译, 北京: 北京大学出版社, 2012, 4.
- [2] Christian, B. *The Alignment Problem: Machine Learning and Human Values*[M]. New York: W. W. Norton & Company, 2020.
- [3] 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): 1357-1358.
- [4] 斯图尔特·罗素. AI新生[M]. 张羿译, 北京: 中信出版社, 2020.
- [5] 尼克·波斯特洛姆. 超级智能[M]. 张体伟、张玉青译, 北京: 中信出版社, 2015.
- [6] Grice, H. P. 'Logic and Conversation'[A], Cole, P., Morgan, J. L. (Eds.) *Syntax and Semantics 3: Speech Acts*[C], New York: Academic Press, 1975, 41-58.
- [7] 刘永谋. AI对齐是一种危险的尝试[N]. 社会科学报, 2024-01-25(06).
- [8] Roche, C., Wall, P. J., Lewis, D. 'Ethics and Diversity in Artificial Intelligence Policies, Strategies and Initiatives'[J]. *AI and Ethics*, 2023, 3(4): 1095-1115.
- [9] Friederich, S. 'Symbiosis, Not Alignment, as the Goal for Liberal Democracies in the Transition to Artificial General Intelligence'[J]. *AI and Ethics*, 2024, 4(2): 315-324.
- [10] Dreyfus, H. L., Dreyfus, S. E. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*[M]. New York: The Free Press, 1986.
- [11] Dennett, D. 'Cognitive Wheels: The Frame Problem of AI'[A], Hookway, C. (Ed.) *Minds, Machines and Evolution*[C], New York: Cambridge University Press, 1984, 129-152.
- [12] Dreyfus, H. L. 'Why Computers Must Have Bodies in Order to Be Intelligent'[J]. *The Review of Metaphysics*, 1967, 21(1): 13-32.
- [13] Dreyfus, H. L. *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*[M]. New York: Oxford University Press, 2014.
- [14] Duan, J., Yu, S., Tan, H. L., et al. 'A Survey of Embodied AI: From Simulators to Research Tasks'[J]. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022, 6(2): 230-244.
- [15] Pfeifer, R., Bongard, J. *How the Body Shapes the Way We Think: A New View of Intelligence*[M]. Cambridge, MA: MIT Press, 2006, 82-85.
- [16] Brooks, R. A. 'Intelligence Without Reason'[A], Mylopoulos, J., Reiter, R. (Eds.) *Proceedings of the 12th International Joint Conference on Artificial Intelligence, Volume 1*[C], San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1991, 569-595.
- [17] Froese, T., Ziemke, T. 'Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind'[J]. *Artificial Intelligence*, 2009, 173(3): 466-500.
- [18] Froese, T., Taguchi, S. 'The Problem of Meaning in AI and Robotics: Still with Us After All These Years'[J]. *Philosophies*, 2019, 4(2): 1-14.
- [19] Ororbias, A., Friston, K. 'Mortal Computation: A Foundation for Biomimetic Intelligence'[EB/OL]. <https://arxiv.org/abs/2311.09589>. 2024-05-18.
- [20] Licklider, J. C. R. 'Man-Computer Symbiosis'[J]. *IRE Transactions on Human Factors in Electronics*, 1960, HFE-1(1): 4-11.
- [21] Gerber, A., Derckx, P., Döppner, D. A., et al. 'Conceptualization of the Human-Machine Symbiosis: A Literature Review'[A/OL], Hawaii International Conference on System Sciences 2020 (HICSS-53)[C], Grand Wailea, Hawaii, 2020, https://aisel.aisnet.org/hicss-53/cl/machines_as_teammates/5. 2024-05-18.
- [22] Hardin, G. 'The Competitive Exclusion Principle'[J]. *Science*, 1960, 131(3409): 1292-1297.
- [23] 阿克塞尔·霍耐特. 为承认而斗争[M]. 胡继华译, 上海: 上海人民出版社, 2005.
- [24] 查尔斯·泰勒. 承认的政治[A], 汪晖、陈燕谷: 文化与公共性[C], 北京: 生活·读书·新知三联书店, 1998, 290-337.
- [25] 丹·扎哈维. 胡塞尔现象学[M]. 李忠伟译, 北京: 商务印书馆, 2022, 149.
- [26] 丹·扎哈维. 现象学入门[M]. 康维阳译, 北京: 商务印书馆, 2023, 120.
- [27] 赵汀阳. 如何定义跨主体性?[J]. 读书, 2023, (5): 3-13.