

# 人工智能伦理漂洗现象及其治理

## The Ethical Washing in AI and Its Governance

杨进 / YANG Jin 杜严勇 / DU Yanyong

(同济大学人文学院, 上海, 200092)  
(School of Humanities, Tongji University, Shanghai, 200092)

**摘要:** 人工智能伦理漂洗现象在现实中的频发引起了学界和公众的重视。从概念上而言, 作为一种行动的伦理漂洗呈现出两种要素特征: 动机上的虚伪性与结果上的虚假性。就其成因而言, 伦理漂洗的形成受到动机因素(生产者的利益追求与消费者和管理者的安全-权益需求)与外部因素(人工智能伦理的有效性, 人工智能对伦理的冲击, 信息差与治理成本差)的综合性决定。由于伦理漂洗在实践与心理层面都造成了负面的伦理影响, 对它的治理需要从内(道德心理)与外(现实制度)两个层面展开。

**关键词:** 人工智能伦理 伦理漂洗 伦理治理

**Abstract:** The widespread existence of the ethical washing in AI has attracted the attention of the academy and the public. Conceptually, ethical washing as an action is characterized by two elements: hypocrisy in motivation and falsity in outcome. In terms of its causes, the formation of ethical washing is determined by a combination of motivational factors (the pursuit of interests of producers and the safety-rights needs of consumers and managers) and external factors (the effectiveness of AI ethics, the impact of AI on ethics, information gap and difficulties of governance). Ethical washing has caused negative ethical effects on both the practical and psychological levels. In this regard, its governance should be carried out from both internal (moral psychology) and external (realistic system).

**Key Words:** AI ethics; Ethical washing; Ethical governance

中图分类号: TP18; B82 DOI: 10.15994/j.1000-0763.2025.01.002 CSTR: 32281.14.jdn.2025.01.002

人工智能中的伦理漂洗 (ethical washing) 现象, 也称伦理漂蓝 (ethical bluwashing) 现象, 是弗洛里迪 (Luciano Floridi) 借鉴生态伦理学中绿色漂洗 (greenwashing) 概念而发明的一个术语。按照弗洛里迪的定义, 伦理漂蓝是绿色漂洗现象的数字版本, 它是一种“对数字工序、产品、服务或其他解决方案的道德价值和益处, 做出未经证实或误导性的声明, 以表现得比数字过程更有数字道德的不当行为”。

<sup>[1]</sup> 但是严格说来, 蓝色这个色彩与数字时代的伦理现象并没有像绿色与生态之间存在那样本质的直觉关联, 所以其他参与讨论的学者在研究中针对这一现象往往更倾向于使用不带颜色词汇的 ethical washing 来加以表达, 因此本文也以“伦理漂洗”作为术语上的翻译。

如今在人工智能实践活动中, 伦理漂洗现象的出现变得愈加频繁。有报道指出, 一些企业会通过购买“虚假伦理”的人工智能产品

**基金项目:** 国家社会科学基金重大项目“人工智能伦理风险防范研究”(项目编号: 20&ZD041)。

**收稿日期:** 2024年3月7日

**作者简介:** 杨进 (1993-) 男, 贵州兴仁人, 同济大学人文学院博士研究生, 研究方向为科技伦理。Email: yangjinnico@sina.com

杜严勇 (1976-) 男, 四川南充人, 同济大学人文学院特聘教授, 研究方向为科技伦理。Email: yanyongdu@tongji.edu.cn

来满足其应该承担的伦理责任,<sup>[2]</sup>此外还有公司一方面推广“AI向善”(AI for Good)计划,另一方面却向政府和企业客户出售监控技术。<sup>[3]</sup>甚至欧盟委员会于2019年提交的“可信任人工智能伦理指南”(Ethics Guidelines for Trustworthy AI)都因在道德红线上的妥协而被参与起草该准则的学者指控为一个“伦理漂洗的案例”。<sup>[4]</sup>因此,伦理漂洗也越来越受到学者和公众的关注,几乎所有对人工智能伦理原则提出质疑和挑战的研究都会提及伦理漂洗现象,以此证明伦理原则在人工智能实践和治理中收效甚微,<sup>[5]</sup>而捍卫或至少在一定程度上承认伦理原则有效性的研究者对漂洗现象的治理提出了若干建议。<sup>[6]</sup>纵观现有的研究,多将重点集中于伦理漂洗的现象分析方面,而对于概念辨析与成因等方面尚缺乏系统的论述。在此,我们将提供一个关于人工智能伦理漂洗现象的综合性研究:它首先将澄清相关的概念,其次分析此现象的成因及其危害,最后尝试提出一些相应的治理意见。

## 一、概念析义及辨明

我们先来尝试解析弗洛里迪对伦理漂洗的定义,以此从中分解出关于伦理漂洗这一概念的关键要素。伦理漂洗首先被定性为一种不当行为(malpractice),即一种为了获取某种利益而破坏规则或法律的行为。我们可将伦理漂洗作为一种不当行为的结构过程描述为:为了解决人工智能产品和服务中的伦理问题P而采取的行动A并不能达成其目标,基于使行动A在表面上达成目标的意图,采取误导性信息将行为A漂洗成行为Aa,但这两者实际上都不是真正能够解决问题P而应当采取的行为。

从这一行动的结构过程可以看到:伦理漂洗的核心步骤在于采取误导性的信息去包装和粉饰行为A,就其动机而言,伦理漂洗所要达到的目的不在于真正解决原本应该在实践中得到实质性关注的伦理问题,而是意图修饰和包装一些行为,使得它们看上去能够在伦理上具备有效性;就其结果而言,伦理漂洗造成了伦

理表象/假象(ethical appearance),即它具有伦理的外表,却并非真正具有伦理的意义和价值。比如伦理漂洗行为最为常见的形式之一就是企业的道德游说活动,有报道指出,谷歌、亚马逊和脸书(Facebook)等公司在2018年以创下支出记录的方式向美国近一半的参议院议员进行了捐款。<sup>[7]</sup>这种游说行为不仅指向立法层面,对学术界也有相当深入的渗透。例如麻省理工学院媒体实验室前主任伊藤穰一(Joichi Ito)的丑闻披露之后,他在硅谷企业与学术界之间的游说合作关系也被揭示出来。<sup>[8]</sup>还有报道指出,由包括谷歌、苹果和微软等科技巨头在内的65家科技公司所组成的游说协会“数字欧洲”在欧盟委员会制定“可信任人工智能伦理指南”的会议上积极推动了对“红线”的删除。<sup>[9]</sup>另一种常见的漂洗行为是,企业虽然在官方层面承认某项伦理责任,但实际上却没相应的行动。比如某著名视频网站公司高层一直表示对于气候变化政策的支持,但网站的算法却在大量推荐关于气候变化的错误信息。<sup>[10]</sup>

于是,对伦理漂洗概念的分析可以得出两个要点:1.就其动机的虚伪性而言,伦理漂洗是不正当和不道德的;2.就其结果的虚假性而言,作为假象的伦理表象也是不道德的。进而言之,结果的虚假性所带来的危害在实践中加深了道德危机的严重程度,而动机的虚伪性则在道德心理学层面对公众的道德认知和情感造成了冲击。这两点在后文的讨论中将会详细展开。

在明确了伦理漂洗在概念上的要素特征之后,我们来看它与绿色漂洗(另译“漂绿”)的异同。绿色漂洗是一种就“公司的环境实践或产品服务的环境效益而误导消费者的行为”。<sup>[11]</sup>从定义上看,绿色漂洗作为一种行为,其形式结构与伦理漂洗基本上是一致的:为了解决环保问题P而采取的行动A并不能达成目标,基于使行动A在表面上达成目标的目的,从而采取误导性信息将行为A漂洗成行为Aa。可见,伦理漂洗和绿色漂洗之间的相同之处就在于,二者皆是以误导性信息进行伦理价值的漂洗,以形成伦理假象从而获取相关利益的不当行为。如此看来,伦理漂洗可谓是绿色漂洗

的数字化版本。

但是，与绿色漂洗相比，人工智能伦理漂洗表现出至少以下三个方面的新特征。其一，人工智能伦理漂洗具有更强的灵活性与隐蔽性。相比于绿色漂洗行为，伦理漂洗更容易与数字伦理选购联系起来。所谓“数字伦理选购”（digital ethics shopping）是指这样一种不当行为，它在各种可供运用的伦理提议中选择或修改一些伦理标准或框架，以此改造一些现存的行为，从而证明这些标准是后验的。它的不当性在于，选购行为没有根据公共的道德标准去改进行动或实施新的行为，而是在一个小市场内部进行有限的选购来采取伦理行动。<sup>[1]</sup>理解这个定义的关键在于：伦理原则的市场化是这个概念的基点，这意味着伦理漂洗与伦理选购的本质关联在于它们共享这同一个基点。在伦理市场化的情况下，所有伦理原则和主张都带有竞争与营销的色彩，这意味着在这个市场中并不存在一个绝对的主导性伦理标准，各种伦理原则之间的相互竞争使得伦理市场呈现出多元化的态势。

因此，“数字伦理选购”这一现象很好地描述了现实中关于人工智能伦理治理的现状，即各种伦理原则被不断地提出，诸种伦理治理方案也层出不穷，一度使得人工智能伦理学市场颇为繁荣。绿色漂洗现象则与此种情况有所不同，因为前者已经拥有了一个基本共识，即环境保护是公认的基础伦理价值。但是在伦理漂洗中，至少就目前而言，还缺乏像环保这样一个坚定的基本共识。尽管也有学者指出了看似多元化的伦理原则实际上具有趋同的情况，<sup>[6]</sup>但不争的事实依旧是：各个原则和方案之间并不能做到互通和统一。因此，科技企业或相关机构可以根据各自的需要进行数字伦理选购，从而使人工智能伦理漂洗具有更强的灵活性与隐蔽性。

其二，人工智能伦理漂洗具有更强的表演性与欺骗性。近些年来，不少科技企业纷纷成立人工智能伦理委员会，表示要加强对人工智能研发的伦理审查与伦理治理，对伦理风险进行前瞻性防范。但是，对于人工智能伦理委员

会具体的运作程序与实际效果，却鲜有报道，人们普遍担心部分科技企业的人工智能伦理委员会已经变成一种表演形式，沦为伦理漂洗的一种工具。前文提及，在欧盟委员会发布的“可信任人工智能伦理指南”中删除了最初草案中提及的伦理红线，具体包括“禁止使用致命性自主武器，禁止使用人工智能进行公民评估与社会评分，以及原则上禁用人们无法理解和控制的人工智能系统”。<sup>[4]</sup>毫无疑问，明确划定类似这样的伦理红线能够有利于更好地防范伦理风险。故而，伦理红线从最初提出到最终删除，使得制定“可信任人工智能伦理指南”的欧盟委员会高级别专家组的工作在某种程度上也成为了一种表演。

其三，人工智能伦理漂洗可能引发更广泛更深远的伦理风险。绿色漂洗产生于环境保护领域，所引发的伦理风险也主要集中于某个区域的生态环境影响与可持续发展等方面。由于人工智能是公认的通用目的技术，其应用范围非常广泛，而且还有明显的扩大趋势，其潜在的用户群体也是极其巨大的。由此，它造成的伦理风险涉及了公众非常关注的伦理问题，包括侵犯个人隐私与尊严、产生偏见歧视与影响社会公正等。人工智能伦理漂洗很可能加强并放大潜在的伦理风险，强化人工智能产品的消极与负面影响，如果不对其进行有效的伦理治理，很可能影响公众接受人工智能产品，进而导致技术进步速度的减缓，对科技企业和科研机构产生不利影响。

## 二、漂洗现象的成因及其危害

学界目前关于人工智能伦理漂洗的研究更多地把关注点放在了现象描述和治理建言之上，而对于它的成因尚缺乏系统和细致的分析。相较之下，以德尔马斯（Magali A. Delmas）和博尔巴诺（Vanessa Cuereil Burbano）的工作为代表的对绿色漂洗的驱动因素研究要更加系统和成熟。因此在对漂洗现象展开成因分析之前，不妨参考关于绿色漂洗现象的成因研究。简言之，他们将绿色漂洗的驱动因素划分为三个方面：外

部、组织与个人,并从管理、政策、社会学和心理学等方面对这些因素的影响情况展开分析。<sup>[11]</sup>但是他们的研究在方法论上基本采取了一种静态结构的方法,并没有对这一行为的动态特征加以强调。而我们之前的分析则重点关注伦理漂洗行为的动态过程,因此我们对于伦理漂洗的成因分析将会从这个行动的各个方面展开。首先需要对这个行动本身的发生过程展开描述,明确各个参与方的相关行为,在此基础上分析它们各自的行为动机,以及对这些动机产生决定性影响的各种外部因素,对二者(动机和外部因素)的综合分析才构成了一个完整的成因探究工作。接下来我们逐步展开:

首先,我们可以对人工智能产业链条上所有与伦理相关的活动进行一般性的刻画,并将其描述为一个综合过程,在其中几种类型的参与者展开了它们互有联系的行动:(1)产品生产方:其中大致可分为两个群体:管理决策层和个体员工,他们合作负责制造产品和提供服务,并保证达到伦理合格;(2)伦理生产者:机构、学者和企业的专家以各种方式提供伦理原则;(3)消费者:购买和使用产品与服务,并关注其伦理合格性;(4)公共管理者:监督和治理产品与服务的伦理问题,并在这一过程中存在寻租的可能。值得注意的是,这个过程不是绝对线性的流程,而是围绕智能产品与服务的伦理合格性展开的一个综合过程,其间的各个要素具有交互性的关系。

基于这些环节中所发生的行为,我们可以就他们各自的动机展开如下分析:(1)产品生产方:为达成伦理合格而同时又必需兼顾成本与利润;(2)伦理生产者:一方面出于工作需要,一方面也受到名利等利益因素驱动;(3)消费者:需要伦理合格性来保证自身权益(安全,隐私等等)不受侵犯;(4)公共管理者:政府或社会需要伦理安全维持公共社会的秩序稳定。我们可以再进一步将这些动机总结为两类:(1)对于两类生产方而言,其动机显然是基于利益原则;(2)对于后两类参与者而言,其动机可以归结为基于安全和权益原则。伦理漂洗现象的发生表明:在利益动机与权益-安全动

机之间发生了冲突,而使得它们不和谐的决定性因素一方面存在于市场化活动中的结构性利益冲突,即利益要求的最大化原则势必对权益和安全产生一定的影响(尽管这一冲突可以通过各种途径加以限制,但其结构性的存在并不因此改变或消失)。

具体而言,当产品生产方在决定选择何种伦理原则以达成伦理合格的时候,出于经济利益的原则,成本节约型或高性价比方案自然更受青睐。尤其在面对由人工智能伦理学原则的多元化与市场化所导致的原则有效性问题这一困境时,对成本的节约需求变得更为急迫。实际上,在伦理有效性不能够得到保证的情况下,生产方对伦理问题的处理将会更加消极,因为既然在采取伦理保障措施的情况仍旧难以保证有效性,那么转而采取伦理漂洗的方法,反倒能在成本上更加节约的同时还能够以伦理假象的方式完成产品的伦理漂洗,使之达到表象上的合格。比如,被微软裁撤的伦理与社会团队成员在接受采访时表示,尽管在企业内部有员工出于伦理考虑对研发工作提出限制,但是公司的领导层出于竞争压力和对效益的追求,对长期性的伦理问题开始失去耐心。<sup>[12]</sup>故而即便在产品生产方内部存在着员工与高层之间在伦理态度上的区别,但是利益原则才是最终的决定性因素。因此从产品生产方而言,造成其进行伦理漂洗的决定因素是人工智能伦理的有效性困境与成本考虑之间的叠加影响。

伦理生产方的利益动机一方面体现为学者群体对名利的追求,<sup>[13]</sup>另一方面也往往受到科技公司对于学术界的影响,这使得学术界的研究结论可能倾向于支持公司的利益。<sup>[14]</sup>虽然我们不能忽视在企业内部的研究人员和员工出于对伦理原则的坚持而进行的斗争,比如微软和Salesforce等科技巨头的员工愈发强烈地反对公司使用人工智能来追踪移民和进行无人机监控等不当行为。<sup>[15]</sup>但是如果考虑到著名的谷歌案例所呈现的现实状况:格布鲁(Timnit Gebru)团队因为拒绝接受谷歌高层对于其研究论文的不公正对待而遭到解雇,<sup>[16]</sup>那么从结果上而言,至少应该承认利益原则的力量深度参与了伦理

生产方的行为。其次,伦理生产方的行为也深受人工智能伦理有效性的影响,因为正是在有效性方面存在的困难,才使得多元化的原则体系能够在理论与实践层面获得立足之地。每一种伦理体系都能够解决一部分问题,这虽然不能满足普遍有效性的要求,但也能够迎合很多实际的需要。再加上数字伦理选购行为在实际上不断地用经验来修正理论,导致人工智能伦理学被论证为后验的,这就更加为多元主义提供了支撑:既然伦理原则在这里是需要经验来证实和修正的,那么就很难认定存在先验的与既定的原则或框架,这就更进一步清除了一元论思维在人工智能伦理学中的影响。<sup>[17]</sup>

在此需要强调的是,我们并没有把伦理漂洗视为产品与伦理生产者的单向度行为,而是认为消费者和管理者并非完全是被动的承受者,他们的行为同样在积极意义上建构了伦理漂洗。因此对他们的动机和影响动机的决定性因素进行考察是讨论伦理漂洗成因的必要环节。就此而言,人工智能伦理的有效性对于消费者和公共管理者而言也同样影响深远。但这种影响更多地体现为一种怀疑的情绪,即他们出于对伦理原则本身之有效性的怀疑,从而更加重视人工智能伦理产品与服务的伦理合格性。另一方面,消费者和管理者对人工智能伦理的重视还受到这样一个更加宏观的因素的影响:人工智能的发展对伦理本身的冲击和影响(其负面的影响尤为显著)。这种冲击造成了一种普遍的对于人工智能的恐慌情绪。尽管不乏对于人工智能持乐观态度者,但是在一定程度上代表社会普遍心理的大众文化中仍旧流行各种危害论与威胁论。这种恐慌情绪对消费者产生的影响体现为,他们对保护信息安全表现出特别的关注。<sup>[18]</sup>而对管理者而言,伦理风险对于社会稳定秩序所带来的冲击同样是对其管理工作的一个巨大挑战,从各国政府和机构对于人工智能伦理治理方案的强烈关注与切实部署来看,这种恐慌情绪是其理性建构活动的一个心理学基础。<sup>[19]</sup>

于是,就消费者和管理者而言,他们一方面怀疑人工智能伦理原则有效性的,一方面恐

慌人工智能伦理风险,这共同造成了他们对于人工智能伦理合格性的严重关切,于是他们必然对此提出更高标准的要求。这种高标准要求对于生产方来说是一个巨大的外部压力,但与此同时,这种伦理要求也反映了伦理市场的需求,正是这种需求的倾向使得伦理漂洗成为可能。因为对伦理的需求一旦以如此明显的方式被表达出来,那么针对这些需求的各种迎合方案便会不可避免地出现。可以说伦理漂洗就是对这种伦理需求的针对性安排,它所造成的伦理表象能够起到实际作用,一方面固然因为其误导性的信息具备迷惑性,但另一方面也是因为它的确满足了各方的伦理需要,安抚了他们对于伦理的迫切心态。而这之所以能够完成,还有赖于信息差和治理成本差这两种外部因素的参与。就掌握的信息而言,生产方势必在技术和专业层面上具有不可比拟的优势,正是利用这种在技术层面的信息优势,使得伦理原则在转换到产品服务中的时候,监管方和消费者其实都很难完全弥补信息差所带来的劣势,于是只能选择相信产品的伦理合格性。因为质疑与验证的成本不仅对于个人而言是难以应付的,而且即便对于公共管理方而言,监管成本的考虑往往也是一个挑战。从这个过程可以看到,在生产方与消费者和管理者之间存在着信息差和治理成本差,这两种差距对造成漂洗现象产生了关键性影响。

综上,我们可以将伦理漂洗的成因划分为以下因素的综合影响:(1)动机方面:生产者的利益追求和消费者与管理者的安全-权益需求;(2)外部因素:人工智能伦理的有效性困境,利益因素,人工智能对伦理的负面影响,信息差与治理成本差的普遍存在。

在厘清伦理漂洗的成因之后,其所产生的各种危害也就能够得到更加深入的讨论。比埃蒂(Elettra Bietti)从道德哲学的立场认为对伦理漂洗的谴责“导致了一种进行伦理抨击(ethical banning)的倾向。这包括对伦理和道德哲学的轻视”。<sup>[13]</sup>因此,为了避免对伦理漂洗的抨击可能导致的对伦理的否定倾向,我们需要基于以治理为导向的理性主义批判态度对

伦理漂洗可能造成的危害进行揭露。

尽管伦理漂洗的成功运作并不代表其效用可以持续存在,但是,伦理漂洗毕竟使得假象起到了作用,这导致的虚假后果使得现实中的伦理风险不但没有得到规避,反而在假象的庇护下继续存在,甚至其危害在持续加深。这主要体现在:一方面,伦理漂洗导致了关键的伦理问题被遮掩,使得伦理风险提升;另一方面,伦理漂洗会造成伪善的流行,它排除了正确的或相较而言更为合理的治理原则与方案。此外,伦理漂洗带来的假象还可能被利用,为继续维护当前的不良制度进行辩护。<sup>[20]</sup>而当伦理漂洗被揭露于现实时,它所造成的危害还将持续。在此,我们特别关注它对于公众的伦理认知和情感所造成的负面影响。具体来说,这种持续性的假象一旦因为各种原因破灭后,其对公众的道德情感和认知会造成两个方面的负面影响:首先,生产方虚伪的道德动机使得公众更加怀疑人工智能伦理原则的有效性和治理方案的可行性;其次,这种不信任态度加剧了对于人工智能的恐慌与否定情绪,这对于整个人工智能事业发展而言都将产生颇为负面的影响。

最后,就相关参与者的层面而言,伦理漂洗所造成的危害则体现为:(1)对于生产者而言,经济效益受损,道德公信力下降,自身的道德形象受损,尤其会加剧消费者不信任态度。<sup>[21]</sup>(2)对于消费者而言,权益和安全都受到侵犯。并且伦理漂洗造成的假象加剧了侵犯的危害程度,比如用户选择相信某种信息服务所做出的隐私保护承诺,从而将原本不会分享的数据上传,这将造成更为严重的隐私侵犯。这种侵犯如果没有被揭露,那么消费者受到的欺骗是更加严重的;而一旦其被揭露,消费者的道德认知和情感将受到打击,从而产生不良的心理情绪。因此伦理漂洗对消费者的危害是全方面的。(3)对于管理者而言,即便它在合法范围内揭露和治理了伦理漂洗,但伦理漂洗的成功还是一定程度上体现了其监管工作的失败,故而其公众形象也可能受到负面影响。更有甚者,如果管理者在其中参与了漂洗活动,那么其公信力将受到严重损害,其腐败行为亦

有违法之嫌。

### 三、人工智能伦理漂洗的治理建议

通过上文对人工智能伦理漂洗在概念、成因和危害方面的讨论,其基本事态在理论层面上得到了展示。由此,我们的讨论将进入到实践层面,即对治理相关问题的分析。在此我们先就已有的研究进行梳理。现有文献对人工智能漂洗治理的研究路径大致有两种:1.在原则上采取较为宏大的哲学和伦理学体系,从理论层面上进行建构;2.相较于第1种路径,更多的学者采取了务实的经验主义态度,提出了较为细致的具体治理方案。

我们先来看第1种路径,其代表性的成果是卡伦(Karen Yeung)等三位学者的研究,他们认为“国际人权标准为人工智能提供了一系列最有希望的道德标准”。针对人工智能伦理漂洗提供的治理方案从原则上而言是一种“以人权为中心的设计、审议和监督”的治理框架。这个框架倡导人工智能在以下四个方面都必须遵循基于人权规范的核心原则:(1)设计和审议;(2)评估与测试;(3)独立监督、调查与制裁;(4)可追溯性与证据。这一框架的治理核心就在于基于对人权风险的评估去衡量人工智能伦理治理在学科、组织、行业和决策等层面的各种工作。<sup>[22]</sup>

相较于这种在宏观原则上的建构,第2种研究路径的工作则提出了更多较为具体的细则性方案。例如,佛罗里迪认为治理的要点在于透明性和教育,主张需要对伦理漂洗相关的所有参与者进行关于“何为有效道德实践”的教育。<sup>[1]</sup>瓦格纳(Ben Wagner)提供了一个相对更加全面的治理方案:(1)尽早确定所有的利益相关者,并定期进行更新;(2)提供在外部上独立的监管机制;(3)确保决策程序保持透明性;(4)制定合理且稳定的标准体系;(5)确保伦理规范不会取代基本权利和人权;(6)明确现有法律和监管框架与伦理规范之间的关系。<sup>[23]</sup>

通过上述文献分析,我们可以看到现有研究所提供的治理方案既有在原则上取法人权理

论的建构，也有在具体措施上提出的现实主义建言。基于本文对伦理漂洗的各种分析，并综合考虑现有研究的成果，我们提出一个综合性的治理方案的建议，它分为内、外两个部分。外部方案主要解决由信息差和成本差这两个现实鸿沟造成的四个问题：

生产方与消费者的信息差：这是信息差距最为显著的鸿沟所在。无论是作为产业实体的生产方还是作为知识实体的机构或学者，都对消费者个体形成了巨大的信息特权。可能缓解这一困境的方案首先需要将重点放在教育上，<sup>[24]</sup>这要求学术机构开展科技伦理的教育。其次，公共管理者和学术机构应该为消费者在获取信息途径上的便利性提供及时的帮助。<sup>[25]</sup>教育的改进与普及将会对公众的道德认知水平有所助益，这对缩减消费者层面的信息差起到了基础性的作用。就生产者方面而言，也有主动消除信息差的义务。在合法保护商业秘密的情况下，产品的生产方应当保持伦理设计和开发的透明性，并主动将相关的有效性测试程序全面地呈现给公众和监管机构；企业方面的研究人员与学者也可以参与到教育与普及的工作中。而学术机构除了教育功能之外，也必须承担向公众宣传相关知识的义务，并在理论层面展开监管和批判的工作。

生产方与监管方的信息差：这里的主要矛盾存在于产品生产方与监管方关于产品与服务的伦理设计及合格性之间的信息差距。处理这种矛盾的主要措施应当把透明性原则放在首位，这要求生产方在合法的范围内对伦理的设计、开发与检测等一系列过程必须保持程序上的透明性，并且要以多元的形式进行展示以确保其全面性。<sup>[26]</sup>监管方在信息透明的情况下，还需要提升相关的伦理认知水平，以确保能够更加公平和正确地评估伦理合格性。目前，国内外人工智能学术团体与科研管理部门都高度重视对人工智能科学研究与产品研发进行伦理评估与伦理审查，并陆续出台了相应的政策与制度。随着伦理评估与伦理审查工作的全面进行与发展成熟，可以在很大程度上降低生产方与监管方的信息差。

生产方与消费者的成本差：消费者与生产者在两个方面存在巨大的成本差，一是在伦理信息的认知与获取方面，二是在伦理的监管和举证上方面。前述关于第一类矛盾的治理建议同样适用于第一个方面的成本差问题，即需要提供便利和及时的教育，并且这种教育除了其学院化的方面外，更应当把学术知识进行公益性质的公共宣传和教育，甚至可以考虑将其作为公民教育的一部分进行推广。在监管和举证方面，消费者需要来自公共机构和政府方面的支持，相关的监督和执法机构必须在这方面保障消费者维护权益的渠道。

生产方与监管方的成本差：生产者与监管者在信息上的成本差可以通过保障透明性原则的相关措施加以缓解，而当我们处理二者在监管上的成本差距时则需要抱持更为现实的态度，因为这里涉及到了核心的矛盾：利益原则与公益原则的冲突。这一冲突如果在一定程度上得到和解，那么需要杜绝寻租行为；而如果它没有得到解决的话，那么就需要采取一个现实主义的态度，通过要求生产方与监管方的共同合作，在法律细节上完善对伦理合格性的检测、评估以及追溯性方面的各项技术性条款，从而有效地缓解矛盾。在实施法律的监管层面上可以引入第三方监管的代表来平衡利益相关者的可能关联。<sup>[27]</sup>

以上提供的治理方案之所以被称为外部的建议，在于它们关注的核心点是从制度层面上解决或缓解相关的具体实践问题。这也是相关研究方案所普遍采取的模式，除此之外，基于本文对伦理漂洗的分析，我们发现其中还存在着对道德情感方面的治理需要。具体而言，对道德情感方面的治理分为预期建设与事后干预：

预期建设：这一阶段的要点是对公众关于人工智能伦理的接受态度进行引导，使之向现实主义和理性主义方向发展。<sup>[28]</sup>这首先要求在教育与舆论方面引导公众对人工智能伦理的有效性采取现实主义态度，并杜绝商业宣传和文化作品中关于人工智能伦理有效性不切实际的、甚至于刻意误导的描述。其次，应该将理性主义作为引导公众伦理态度的基本原则，

这要求在公共层面上展开真实有效的伦理认知教育,培育理性和科学的精神。再次,在人工智能产品研发过程中,通过建构性技术评估等手段,使所有的利益相关者都参与到技术评估与研发过程当中,使技术研发更好地符合利益相关者的价值观与现实需求。

事后干预:在伦理漂洗被揭露之后,事实性的伦理伤害在个人层面和公共层面都会造成严重的心理危害。此时的干预一方面需要考虑个人在被侵犯后的心理干预与治疗,社会方面需要提供切实的心理专业援助。另一方面,公众层面的怀疑论态度也需要得到有效的应对,相关部门应该对遭到破坏的道德信心进行重建和维持。

与外部方案不同的是,内部方案考虑的问题虽然也是从制度层面上提供建议,但其关注点不在于经济利益与法律正义等现实主义方面,而是在于虽然抽象但却真实影响人类生活的道德情感方面。本文建言的立场是,伦理漂洗所带来的伦理风险在现实层面的危害与其在道德心理上的危害同样需要得到治理层面的关切。这种内外一体的方案框架可以为治理伦理漂洗提供一组更为全面的建议。

最后,我们想特别强调一下人工智能伦理研究者在治理中应该发挥更为重要的作用。在制定“可信任人工智能伦理指南”的专家组成员中,只有4位伦理学家,另外48位专家主要来自科研机构与产业界。在这样一个主要由科研人员与企业家构成的专家团队中,伦理学家基本上没有多少话语权。来自科研机构与产业界的专家对于主导人工智能技术研发方向的重要性是不言而喻的,但他们不能主导人工智能的伦理评估与社会治理。<sup>[29]</sup>人工智能伦理研究者应该在理论探讨的基础上,针对人工智能伦理漂洗等社会现实问题,并与科研人员和监管部门密切合作,提出具有针对性与可操作性的解决措施,切实防范人工智能伦理风险。

## 结 论

通过本文的工作,我们可以提供一个对人

工智能伦理漂洗的整体概观:(1)通过在定义上的要素分析,伦理漂洗作为一种行为,其主要特征体现为:动机上的虚伪性和结果上的虚假性。(2)伦理漂洗的形成受到动机因素和外部因素的综合性决定。(3)伦理漂洗在现实与心理方面的危害为其治理方案的建议提供了具体的指引。可以看到,人工智能伦理漂洗是一个综合性的现象,对它的分析与治理必须从各个层面出发,采取理性的现实主义立场,并认识到它存在的长期性及其对现实影响可能发生的变化。所以我们不能期待一劳永逸地解决人工智能伦理漂洗,它对于人工智能伦理学而言是一个值得长期关注的话题。

## [参 考 文 献]

- [1] Floridi, L. 'Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical'[J]. *Philosophy & Technology*, 2019, 32(2): 185-193.
- [2] Simmonds, K., Paul, A., Gibbs, J., et al. 'Ethical Measurement in AI and How to Avoid Ethics-Washing'[EB/OL]. <https://www.wombledonnickinson.com/uk/insights/articles-and-briefings/reconnect-ethical-measurement-ai-and-how-avoid-ethics-washing>. 2023-04-26.
- [3] Johnson, K. 'How AI Companies Can Avoid Ethics Washing'[EB/OL]. <https://venturebeat.com/ai/how-ai-companies-can-avoid-ethics-washing/>. 2019-02-17.
- [4] Metzinger, T. 'EU Guidelines: Ethics Washing Made in Europe'[EB/OL]. <https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html>. 2019-04-08.
- [5] Munn, L. 'The Uselessness of AI Ethics'[J]. *AI and Ethics*, 2022, 3(3): 869-877.
- [6] 吴红、杜严勇. 人工智能伦理原则:从原则到行动[J]. *自然辩证法研究*, 2021, 37(4): 49-54.
- [7] Zubof, S. 'The Coup We are not Talking About'[EB/OL]. <https://www.nytimes.com/opinion/sunday/facebook-surveillance-society-technology.html>. 2021-01-29.
- [8] Ochigame, R. 'The Invention of Ethical: AI How Big Tech Manipulates Academia to Avoid Regulation'[EB/OL]. <https://theintercept.com/mit-ethical-ai-artificial-intelligence/>. 2019-12-02.
- [9] Klöver, C., Fanta, A. 'Keine Roten Linien: Industrie Entschärft Ethik-Leitlinien für Künstliche Intelligenz'[EB/OL]. <https://netzpolitik.org/2019/keine-roten-linien->

- industrie-entschaerft-ethik-leitlinien-fuer-kuenstliche-intelligenz/. 2019-08-04.
- [10] Corbin, K. 'Lawmaker Wants to Know Why Climate Misinformation is Rampant on YouTube' [EB/OL]. <https://www.forbes.com/sites/kennethcorbin/lawmaker-wants-to-know-why-climate-misinformation-is-rampant-on-youtube/?sh=27675c623af9>. 2020-01-28.
- [11] Delmas, M., Burbano, V. C. 'The Drivers of Greenwashing' [J]. *California Management Review*, 2011, 54(1): 64-87.
- [12] Schiffer, Z., Newton, C. 'Microsoft Lays off Team that Taught Employees How to Make AI Tools Responsibly' [EB/OL]. <https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs>. 2023-03-13.
- [13] Bietti, E. 'From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics' [J]. *Journal of Social Computing*, 2021, 2(3): 210-219.
- [14] Whittaker, M. 'The Steep Cost of Capture' [J]. *Interactions*, 2021, 28(6): 50-55.
- [15] Hao, K. 'In 2020, Let's Stop AI Ethics-Washing and Actually Do Something' [EB/OL]. <https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/>. 2019-12-27.
- [16] Harris, J. 'There Was All Sorts of Toxic Behaviour: Timnit Gebru on Her Sacking by Google, AI's Dangers and Big Tech's Biases' [EB/OL]. [https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases?CMP=share\\_btn\\_url](https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases?CMP=share_btn_url). 2023-05-22.
- [17] Mittelstadt, B. 'Principles Alone Cannot Guarantee Ethical AI' [J]. *Nature Machine Intelligence*, 2019, (1): 501-507.
- [18] Obar, J. A., Oeldorf-Hirsch, A. 'The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services' [J]. *Information, Communication & Society*, 2020, 23(1): 128-147.
- [19] Rozen, C. 'AI Leaders are Calling for More Regulation of the Tech. Here's What that May Mean in the US' [EB/OL]. [https://www.washingtonpost.com/business/2023/05/31/regulate-ai-here-s-whatthat-might-mean-in-the-us/770b9208-fd0-11ed-9eb0-6c94dcb16fcf\\_story.html](https://www.washingtonpost.com/business/2023/05/31/regulate-ai-here-s-whatthat-might-mean-in-the-us/770b9208-fd0-11ed-9eb0-6c94dcb16fcf_story.html). 2023-05-31.
- [20] Kaspersen, A., Wallach, W. 'Why are We Failing at the Ethics of AI?' [EB/OL]. <https://www.carnegiecouncil.org/media/article/why-are-we-failing-at-the-ethics-of-ai-2021-11-10>.
- [21] Peukert, C., Kloker, S. 'Trustworthy AI: How Ethicswashing Undermines Consumer Trust' [J/OL] [https://doi.org/10.30844/wi\\_2020\\_j11-peukert](https://doi.org/10.30844/wi_2020_j11-peukert). 2020-03-09.
- [22] Yeung, K., Howes, A., Pogrebna, G. 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing' [A], Dubber, D., Pasquale, F., Das, S. (Eds.) *The Oxford Handbook of AI Ethics* [C], New York: Oxford University Press, 2019, 77-106.
- [23] Wagner, B. 'Ethics as an Escape from Regulation. From Ethics-Washing to Ethics-Shopping' [A], Bayamlioglu, E., Baraliuc, I., Janssens, L., et al. (Eds.) *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* [C], Amsterdam: Amsterdam University Press, 2018, 84-89.
- [24] Borenstein, J., Howard, A., 'Emerging Challenges in AI and the Need for AI Ethics Education' [J]. *AI and Ethics*, 2021, 1(1): 61-65.
- [25] Papazolgo, A. 'Silicon Valley's Secret Philosophers Should Share Their Work' [EB/OL]. <https://perma.cc/6KZR-ASJ9>. 2019-08-28.
- [26] Balasubramaniam, N., Kauppinen, M., Rannisto, A., et al. 'Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements' [J]. *Information and Software Technology*, 2023, 159: 107197.
- [27] Nahmias, Y., Perel, M. 'The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations' [J]. *Harvard Journal on Legislation*, 2021, 58(1): 145-194.
- [28] 何勤、朱晓妹. 人工智能焦虑的成因、机理与对策 [J]. 现代传播 (中国传媒大学学报), 2021, 43 (2): 24-29.
- [29] Benkler, Y. 'Don't Let Industry Write the Rules for AI' [J]. *Nature*, 2019, 569(7755): 161-161.