人工智能时代的道德迷思与解蔽

冯永刚 臧琰琰

〔内容摘要〕人工智能时代的美好生活不能离开道德。道德既是人工智能理性的整合器、人工智能发展的导引器，也是人工智能犯罪的抑制器。人工智能在让人类享受技术发展带来的普惠性红利的同时，也引发了人类价值迷失、道德秩序紊乱、道德责任消弭、道德自由衰落、道德信任缺失等一系列道德迷思。为此，要通过强化人对自身存在意义的认知、培养人的人文关怀意识、提高道德推理能力来捍卫人的道德主体地位，使人类成为智能时代道德的运用者和审视者；要设定人工智能研发和使用的道德圭臬，促使价值理性回归，实现工具理性和价值理性的平衡；要通过优化算法设计、规范算法使用、强化算法监督等多种途径构建完善的算法法律法规体系，确保算法合法、公正、透明，从而有效维护智能时代的道德秩序；要将道德原则嵌入智能机器设计，赋予人工智能道德感，从“关系转向”的视角重塑和谐共融的新型人机关系，进而形成人工智能发展和人类福祉增进的双赢局面。

[关键词]人工智能；道德困境；道德责任；人机关系

[作者简介]冯永刚，山东师范大学教育学部教授；臧琰琰，山东师范大学教育学部博士研究生

人工智能是利用计算机的深层算法，通过模拟人类的神经网络对数据库开展深度学习，进而能够像人类那样进行分析、判断和决策的实体或程序。牛津大学信息哲学教授卢西亚诺·弗洛里迪(Luciano Floridi)认为，人工智能是继哥白尼革命、达尔文革命、神经科学革命之后人类自我认知的“第四次革命”。目前，人工智能时代已然到来，人工智能技术已渗透并广泛应用于智能制造、商业服务、教育娱乐、医疗护理、智慧家居等领域，给人类的生产生活、精神文化带来巨大影响。尤其是2022年OpenAI发布的超1750亿参数量的人工智能通用大模型ChatGPT(Chat Generative Pre-trained Transformer)火速出圈，“这是继推出深度学习技术后又一里程碑式的技术革命”。生成式人工智能产品呈现井喷式发展态势，OpenAI的GPT-4、Anthropic的Claude2、百度的“文心一言”、阿里巴巴的“通义千问”等大模型相继问世，文本创作、代码编辑、方案设计、作业辅导等功能都被囊括其中，并且在很多专业测试中，生成式人工智能产品的表现甚至已经优于人类。在与医疗、工业、教育等行业的融合共生中，生成式人工智能正改变着人类社会的生产生活方式，《2023年中国生成式AI企业应用研究》预测，2035年中国约85%的企业将采用生成式人工智能。当前，AI大模型正向通用人工智能迈进，并对人类道德产生前所未有的冲击。我们要充分认识人工智能时代道德的地位和价值，厘清人工智能发展使人类面临的道德难题。人类唯有找准自己的角色，努力寻求道德困境的破解之道，才能创建人工智能时代的美好生活。

一、人工智能时代道德的地位和价值

道德是人固有的内在价值尺度，对于调节人的思想观念、行为方式和人际关系具有十分重要的作用。快速发展的人工智能技术不仅不能代替道德，反而更加依赖道德。只有在道德的调节、规约和指引下，人类才能使人工智能保持工具理性和价值理性的均衡，使之始终朝着促进人类福祉的方向发展，降低智能犯罪的潜在风险。

1.道德是人工智能理性的整合器

马克斯韦伯(Max Weber)认为，现代社会存在工具理性和价值理性两种理性逻辑，人工智能亦同时具有这两种理性。其中，人工智能的工具理性强调其对于既定目标实现的经济性和有效性，而价值理性强调其研发及应用中的合目的性和合道德性。目前，人工智能技术日趋成熟，被广泛应用于工业、商业、教育、医疗等领域，表现出了推动经济社会发展的革命性力量。在这一过程中，人工智能的工具理性愈发膨胀，不断侵蚀着价值理性，极有可能成为人工智能时代社会的宰制力量。人工智能的存在价值和终极目的是促进人类发展、增进人类福祉。如果人工智能的发展只重视工具理性，而忽视人自身的价值，就可能出现很多为了达到目的而僭越道德规范、破坏道德秩序的行为，让人类社会面临前所未有的道德困境和巨大的潜在风险。因此，人类需要对人工智能两种理性的关系予以认真审视，促使人工智能的发展始终指向其存在的终极目的。

人工智能是人类智慧的结晶，人类是人工智能的操控者。所以，当前人工智能价值理性的失衡是人类自身导致的。实现人工智能价值理性的复归，前提是人要做出相应的改变。道德是人固有的内在尺度。正是因为有道德的属性，人类才能不断地自我反思，才能不断修正、调整自己的价值观念和行为方式。通过道德的自省，人类才能主动逃离被人工智能异化的风险，发挥道德实践主体的力量，优化人工智能的研发和应用的全过程，使其为人类的发展服务。而且，价值理性和工具理性并不是非此即彼的对立关系，二者相涉相融、难以分割。所以，人类要充分发挥道德的匡正作用，防止人工智能工具理性误入歧途，促进其价值理性的回归，充分整合人工智能的工具理性和价值理性，使人工智能成为智性和德性交融的统一体。

2.道德是人工智能发展的导引器

从属性看，技术和道德的范畴不同，前者属于事实问题，后者属于价值问题。所以，有人认为技术是中立的，与道德无关。然而，这只是从本体论角度得出的结论。如果从认识论的角度看，技术的研究应用属于人的实践活动。既然实践的主体是人，纵使技术本身隶属于科学范畴，其应用过程也必然蕴含实践主体的目的和价值诉求。诚如美国科学哲学家希拉里·普特南(Hilary Putnam)所说，“每一个事实都有价值负载，每一个价值也都负载事实”。人工智能技术的研发和应用过程也是人践行某种价值理念的过程。所以，道德是能够对人工智能的发展产生影响的。

“道德的基础不是对个人幸福的追求，而是对整体的幸福，即对部落、民族、阶级、人类的幸福的追求。”所以，无论什么时候人类都不能置他人、社会的利益于不顾，而一味追求自己的个人幸福，这不仅是道德是否允许的问题，而且是由人的存在状况决定的是否可能的问题。道德可以从“真”“善”“美”三个维度对人工智能的发展予以引领，确保人工智能使用目的的合理性和使用途径的正当性，为其注入能够实现长久理性发展的生命动力。同时，人工智能技术自身的特性决定了必须以道德进行规约。目前的智能机器已经具有了一定的分析、判断和行动的自主性，而且未来还将朝着通用人工智能和超人工智能的方向发展，如若没有道德的规约，任其发展，很有可能会引发技术滥用，导致隐私泄露、伦理失序、社会不公等诸多问题，届时人工智能可能不仅不会给人类带来普惠的红利，反而会将人类置于被人工智能反噬的危险境地。

3.道德是人工智能犯罪的抑制器

在人工智能时代，智能机器大量涌入人类日常生活，既给人们带来极大便利，也带来日益增大的风险。近年来，智能机器伤人或人类利用智能机器实施犯罪的事件越来越多。前者是机器本身的设计缺陷导致的，随着技术的日益成熟或能得到解决；而后者是人有意实施的犯罪行为，和技术本身无关，故不能通过发展技术杜绝。由于人工智能犯罪有别于传统犯罪形式，在主体、罪过和行为的认定方面存在很大困难，而且智能机器未来发展存在太多的不确定性，当前还不能以明确的法律条文对人们利用人工智能的行为做出强制性规定。种种原因导致目前针对人工智能犯罪的立法还处在探索阶段，但是人工智能犯罪已经出现。2017年9月，浙江省绍兴市出现了全国首例利用人工智能实施犯罪的案件，犯罪分子利用智能机器来识别验证码，窃取公民个人信息，进而实施网络诈骗。而且，针对人工智能犯罪的立法流程较长，而人工智能犯罪手段的更新速度却非常快，所以用现有的法律手段很难达到遏制人工智能犯罪的目的。在此情况下，道德对法律的弥补作用显得尤为重要。

人是共生性的存在。即使在人工智能时代，人只要在社会中生存就需要与他人建立各种关系，只是关系的存在空间和形态发生了变化。随着交往空间向虚拟世界延伸，人们不仅和身边的人建立关系，而且可能和另一个空间的素未谋面的人建立关系。然而，不管是哪个空间的关系都可能受利益驱动，利益是开展交往活动的前提。而道德超越了法律的覆盖范围，其调整领域囊括了一切社会关系。道德是人固有的内在尺度，如同人们内心的一杆秤，可帮助人们更加理性地认识人与人之间的利益关系，并自主地调节行动，使自己的欲求趋于理性；同时它还能影响社会舆论，借由舆论的感召力引导人们提高思想觉悟，进而自觉地规范行为。在人工智能时代，道德的价值就在于让人审视自己使用人工智能的方式和行为，看其是否符合道德标准，是否侵犯他人的权益，然后主动调节自己的行为，从而为构建智能时代安全稳定的社会秩序发挥积极的促进作用。

二、人工智能时代的道德迷思

生活在人工智能时代的人们，虽然享受到了智能技术发展带来的物质方面的空前红利，但同时面临着与技术化生存相伴而来的前所未有的道德问题。早在100多年前，马克思就警示道：“在我们这个时代，每一种事物好像都包含有自己的反面。我们看到，机器具有减少人类劳动和使劳动更有成效的神奇力量，然而却引起了饥饿和过度的疲劳。……技术的胜利，似乎是以道德的败坏为代价换来的。”智能机器全面融入人类生活，可能使人类产生价值迷失，弱化人类的道德责任和道德自由，进而影响人与人之间道德信任的建立，对整个社会道德秩序的稳定带来不利影响。

1.智能神话催生价值迷失和价值困惑

进入人工智能时代后，在工作领域，很多职业被智能机器所取代，给人类的主体地位、存在价值带来巨大冲击；在生活领域，人们陷于各种智能软件构建的虚拟空间中难以自拔，人工智能创造的智能神话和泛滥的非主流信息，使价值迷失和价值困惑日渐萌生。

其一，智能机器代替人类劳动令人类产生价值危机。由于智能机器在数据的接收、运算等方面的能力远高于人类，可不知疲倦地准确地完成任务指令，且长期运行成本要远低于人力成本，所以越来越多的职业正在被智能机器代替。虽然人类从沉重、烦琐的劳动中解放了出来，有了更多的时间和精力去追寻自己的兴趣，但是这也弱化了人自身之于国家和社会的意义。人的价值是在劳作中体现出来的。“人一旦失去操劳，在很大程度上也就失去了意义赖以产生的大地”。除了少部分智能机器的研发设计人员和高层管理人员，大多数人都可能沦为“无用阶级”。于是，有人在心底会滋生出一种“无用感”开始自我否定，对自己存在的价值和意义产生怀疑。虽然人的身体得到了解放，但是心灵背负了枷锁。在这样的生存状态下，即便物质生活再丰富，人的精神世界也是贫瘠的。在人工智能时代，人类面临前所未有的价值危机。

其二，智能平台低俗信息泛滥令人类产生价值困惑。在人工智能时代，每个人都可成为信息的生产者和传播者。很多网红、“大V”为了赚取流量、吸粉，刻意传播负面新闻，发布恶俗视频。智能平台的信息筛查机制还不健全，对有些信息的合法合规性难以判断，导致很多有悖于社会主流道德观念的信息也被上传至网络。在虚拟的网络空间中，高尚道德、底线道德和违背道德的信息出现的频率和优先级并非由道德层级的高低决定，而是由赋值市场的竞争结果决定。一如“坏消息法则”所指出的那样，负面、低俗信息的传播速度远远超过体现社会主流道德观念的信息，主流价值的渗透和教化作用远远无法消除恶俗负面信息带来的消极影响。如今，刷微博、快手、抖音等已经成为很多人打发闲暇时光的主要方式。在智能平台上形形色色的信息、短视频的长期影响下，人们内心自觉或不自觉地滋生出焦虑感和虚无感，并在人生观念、道德认知等方面产生很多困惑，对于人生的追求、活着的意义、道德价值等问题找不到明确的答案，出现了智能时代的价值迷失。

2.算法偏见妨碍道德秩序的稳定

算法是一种逐步解决问题或完成任务的方法，是人工智能的核心。人工智能时代任何问题的解决都可以归结为算法。在这个由数据构成的世界里，人类变得越来越依赖算法。然而，算法并不是价值无涉的。受政治内嵌和资本介入的影响，算法推荐可能会加速负面低俗信息传播，固化人们狭隘的道德认知。算法决策可能得出带有歧视、偏见的不公论断，侵犯特定群体的切身利益。在道德认知狭隘化导致道德共识难以形成和特定群体因受到不公对待而产生抵抗情绪的双重作用下，社会道德秩序的稳定也受到影响。

其一，算法推荐阻碍了道德共识的形成。道德共识是不同行为主体达成的对于道德生活的应然状态的共同认识。道德共识的形成对于个体精神生活的平静和整个社会道德秩序的稳定都至关重要。如今，为了提高流量，很多数字媒体类手机应用在信息推送时都选择基于用户画像的个性化算法推荐机制。根据用户的个人数据和浏览习惯为其建立“数字档案”，向用户推送与其兴趣相符的信息，最大程度满足用户的阅读期待。然而，算法推荐忽视了推送信息本身的道德价值，导致有的用户深陷低俗、拜金等挑战道德底线的信息构成的封闭媒介空间中，在浏览同类重复信息的过程中不断强化自己固有的道德认知，削弱了批判性思维的能力，失去了走出信息牢笼、与他人建立道德共识的机会，影响了整个社会范围内道德共识的达成。

其二，算法决策损害了社会的公平正义。算法具有强大的计算能力且更加精确，能够排除情绪干扰和主观偏见，致使很多人错误地认为，同人类决策相比，算法决策更加公平公正，故而提倡在教育、医疗、司法等很多领域以算法决策取代人类决策。然而，大多数人对算法的运行并不了解，算法并非如人们想的那样公正。算法以数据为基础，人工智能基于网络上已有资源进行数据收集，这些数据体现着发布者的道德准则、价值观念，其中可能会存在性别歧视、种族歧视、阶层歧视等问题。一旦这些数据被收入程序代码，就会导致算法决策对特定群体产生偏见和歧视。如美国目前使用的辅助量刑和罪犯再犯风险评估的COMPAS等软件在算法决策中对有色人种有严重的种族歧视"’o歧视现象隐藏于人工智能算法“公正”的外衣下，使得社会公正日趋失调，加深了人工智能时代的社会危机，成为人类不能忽视的重要问题。

3.角色替代肇致道德责任的消弭

责任是行为主体对在特定社会关系中的义务的承担，道德责任是人们在一定社会关系中对自己行为产生的后果在道义上的相应承担[12]。人类的社会属性决定了其对国家、社会、集体和他人负有不可推卸的道德责任。然而，伴随着智能机器使用范围的扩大和能力的提高，人类越来越多地把自身角色规定的任务交由智能机器来完成，随即引发了很多值得思考的道德问题。

其一，智能机器是否具备代替人类承担道德责任的能力。早在20世纪初，汉斯•约纳斯(Hans Jonas)、乔尔•费因伯格(Joel Feinberg)等人就曾对科学技术给人类社会带来的正反两方面的影响进行了反思。他们提出承担责任需要满足三个必要条件:一是自身行为给他人和社会带来影响;二是行为的产生受行为主体的控制；三是行为主体能够在一定程度上预见行为

的后果。虽然现阶段的智能机器已能够模拟人类的情感思维、神经机制和认知行为，具有了自主应对道德问题的能力，但有学者认为，人工智能体只有弱自主性，它们即使具备道德能力也并不能理解道德行为和道德准则的关系，不能预见行为导致的后果。按照约纳斯的责任判定标准，人工智能体仍然不能独立承担道德责任。然而，也有学者认为，人工智能体具有显著的自动化特征，可以承担与其智能水平和自治能力相适应的道德责任，如果将其排除在道德责任之外，可能会造成“责任空场”。可见，对于人工智能体究竟能否承担道德责任，学界尚有争议。因此，当人类将自己的角色任务转交给智能机器时，一旦出现责任事故，如何判断责任归属就成为一大问题。

其二，人工智能体的过度使用弱化了人的道德责任。人类社会中的每个人都不是孤立的，只要在社会中生活就要和其他组织、个人建立各种关系，扮演各种角色并承担相应的道德责任。在人履行责任的过程中，各种关系不断得到维系和巩固。而人往往身处于不同的群体中，是多个角色的集合体，需承担多重道德责任。尽管这些责任在很大程度上束缚了人类，使人不能率性而为，阻碍了自身的“逍遥”，但这些责任是人之为人所不能推卸的。正是人身上担负的责任促使人努力进取、积极向上，让人在履职尽责的过程中提升个体人格、彰显社会价值。然而，随着智能机器功能的日趋完善，人们越来越多地将自己的责任转交给智能机器，如人工智能医生可对病人的病情进行快速诊断，人工智能教师可对学生进行精准指导，人工智能保姆可对空巢老人给予照顾，等等。的确，智能机器可以代替我们完成我们的社会角色规定的很多任务。但是，仍有很多任务是智能机器无法完成的。因为智能机器是“冰冷”的，当我们完全把自己的道德责任让渡给智能机器时，即使工作被完成得再好，我们也不能达到他人和社会对我们角色的期待。如果师生间的关爱、医患间的信任、亲人间的温情不存在了，那么人的道德责任也将因智能机器的过度使用而被消解。

4.技术侵入引发道德自由的让渡

自由是人的本质属性。只有仅因自身本性的必然性而存在，行为仅由自身决定的东西才可被称为自由的；反之，如果一物的存在及行为是由他物所决定的，该物便不能被称为自由的。而道德自由是人们不受干涉地做出道德选择和决定的能力。人工智能时代技术的侵入在很大程度上源于人类将自己的道德自由让渡给智能机器。

其一，智能机器干预人的道德行为。道德行为产生的过程是人为自己立法的过程。在信奉某种道德原则并生成内部动机后，人便会基于所信奉的道德原则自主做出“应该做什么”“不应该做什么”的选择。只有发自内心地认为某种行为本身是正确的、应该做的，做出这种行为的人才是真正有道德的人；如果是因为渴望或惧怕行为的结果而被动做出行为，即使行为产生的结果是好的，该行为主体也不能被看作有道德的。而在智能化的生存空间中，智能机器可在人毫无察觉的情况下收集到人的身心特征、社交生活、兴趣爱好、经济状况等各方面的数据，人所有的行为几乎都能被记录下来。这就意味着在行为发生之前，人就已经意识到了自己的行

为会被监控、该行为的发生会引发什么样的结果。所以，在人工智能的促逼下，人为了得到或避免预知的结果可能做出违背主观意愿的行为。

其二，智能机器限制人的意志自由。朱利安•萨沃斯库(Julian Savulescu)把人工智能看作继教育-宗教、生物医学之后增强人类道德的第三种方式。而且，其作用机制不同于教育-宗教，不需要长期的影响和渗透才能对人的道德发展发挥作用；其作用机制也不同于生物医学，不需要运用药物对人的身体和精神进行控制。智能机器被植入道德建议的程序后，不仅能帮助人们进行道德反思，向人们提供道德建议，指导人们的行为，而且具有快速增强人类道德和免于药物副作用的双重优势，因此受到很多人的追捧。但是也有很多学者对利用人工智能来增强人类道德的方式提出了质疑。且不论人工智能“变异”的潜在危险和被植入的道德建议程序是否可行，单就人类依靠智能机器进行道德判断、做出道德选择这一点来看，这实则是对人类意志自由的侵犯。当智能机器监控人的思想和欲望，从外部对人的思想是否道德进行审查，并强制性地让人远离非道德的行为时，即便产生的结果是好的，也是以人牺牲本质性的自由权利为代价的，人的道德自由意志就被智能机器绑架了。

5.虚拟交往阻抑道德信任的建立

道德信任是交往双方对彼此道德素质和行为能力的心理期待，是对对方可否值得依赖的认知判断。道德信任不仅可以激发个体完善的内在动力，对于促进群己和谐、增进普遍社会信任也具有重要意义。进入人工智能时代后，智能机器的大量使用改变了传统的交往主体，降低了交往关系的透明度，数据大量采集导致的个人隐私“前台化”趋势则增加了科技犯罪的可能性，使人们对潜在风险变得更加敏感，从而阻抑了交往双方道德信任的建立。

其一，交往主体的复杂改变了道德信任的载体。交往是道德信任存在的载体。在传统社会中，交往是在物理环境中的人际、群体间和组织间的交往。进入人工智能时代后，智能机器全面融入人类生活，并充当起生产者、陪护者、教育者等各种社会角色。数字化表征的虚拟人也频繁地出现在大众视野中。人工智能技术的发展使虚拟人形态、表情乃至声音，都与真人越来越相似，虚拟人的应用正向传媒、娱乐、医疗、教育、养老等领域快速延伸。真实人类和智能机器、虚拟人正在形成崭新的交往关系，即交往关系突破了真实物理环境的时空界限和人与人的对象界限，已经拓展到现实和虚拟融合的时空，在真实和虚拟的流动空间中，人、智能机器、虚拟人可共同参与建立复合性的交往关系。人工智能时代的交往较之前有了更多的不确定性，既表现为交往关系的偶发性和无感知性，也表现为交往主体间部分责任和义务的难以明确。这意味着人工智能从根本上颠覆了原有的交往关系，在传统和当下的伦理框架下出现了很多无法解决的道德难题，例如在新型复合性的交往关系中道德信任难以建立。

其二，智能犯罪的发生引发了道德信任的危机。人工智能时代，虚拟世界成为重要的生活空间。在虚拟世界中去躯体化、数字化的存在方式使人可以脱离现实世界的参照去构造多种不同的虚拟身份，强大的真实感和沉浸感使人无法摆脱对虚拟世界的依赖，给人的精神和心理状

态带来严重影响，使人们在真实世界的思想和行为方式也发生改变。身份隐匿性使人产生即便违法也不易被发现的侥幸心理，增加了僭越道德、法律红线的概率。以“虚拟”符号身份和“不在场”的方式与对方隔空交往，由于感受不到对方作为活生生的人的反应，犯罪者负罪感减轻了，他们即使给对方造成身心伤害，也可能增加再次犯罪的可能性。另外，人工智能时代各种智能设备的正常运转多以海量数据采集为基础，其中涉及大量的个人隐私数据，如人脸、指纹、虹膜等身份信息以及活动轨迹、消费记录等。经智能数据分析系统分析后，看似无关的数据就会相互联系，反映出被分析对象的性格、价值欲求、兴趣爱好等方面的特征，使智能设备甚至比对象自己还了解他。一旦这些数据泄露，犯罪分子就可能乘虚而入，给人们的生命和财产安全造成威胁。多重虚拟身份和过度的数据采集，加大了智能犯罪的风险，进而也加剧了道德信任的危机。

三、人工智能时代道德迷思之解蔽

人工智能技术将人类带入了一个新的技术化生存的时代。然而，技术化生存和人类的福祉并非总是正向的关系。为了使人工智能和人类社会始终能彼此适应、相互促进，我们一方面要通过道德教育帮助教育对象强化道德责任，提高德性修养，使其能更好地应对智能机器给人类社会带来的伦理道德挑战，另一方面要完善智能时代的伦理原则，规约人工智能的发展，构建和谐共生的新型人机关系。

1.珍视人性：捍卫人的道德主体地位

鉴于人是智能机器的研发者、拥有者和使用者，要想走出道德困境、规避风险，最为关键的是促进人的改变，通过教育使人充分认识人类的道德主体地位，珍视人类在德性、情感等方面的独有优势，不断提高自身的道德判断力，主动承担起自己的道德责任，自觉遵守人工智能时代的道德伦理规范，成为智能时代道德的运用者和审视者。

第一，要强化人对自身存在意义的认知。个体要想摆脱人工智能引发的自我价值危机，唯有依靠自己。只有主观认知改变了，人才会主动走出使人沉沦的物欲世界，去探寻人生的价值意义。而道德教育以促使人德性发展并收获美好生活为终极目的，它直接作用于人的思想，是促使人思想改变的最有效方式。因此，要通过道德教育使人充分认识到人类自身的优势。智能机器的所有行为都由人的逻辑指令引发，它们自身既没有自由意志，也没有任何情绪体验。而人的所有行为都是自我选择的结果，并且人在行动的过程中会产生丰富的情感体验。人可以感受到领导的期许、同伴的鼓舞，可以体会到胜利时的欣喜、失败时的颓丧。至少在可见的未来，人工智能难以拥有这些感性要素，也无法从人类身上剥夺这些要素。另外，人的组织协调能力、审美创造能力等也是优于人工智能的。所以，在人工智能时代，人不能总是盯着自己的弱势，而应该努力去挖掘并展示自己的优势，彰显道德主体地位的能动性，这样才能确证自身价值，彰显人之存在的意义。

第二，要培养人的人文关怀意识。智能化手段和信息技术的应用，在为人类搭建虚拟活动空间的同时，也拉开了真实世界中人与人之间的距离。日常生活中，面对面坐在一起的两个人各自低头看着手机、沉默不语的景象已屡见不鲜。越来越多的人将自己本应承担的责任如对子女的陪伴、对老人的照顾交由智能机器人代劳。人与人之间的关怀和责任正在被弱化、消解。但是，人是共生性的存在，彼此间具有以生命存在为基础的道德联结。人正是通过不断发展关怀、责任等道德属性来实现自我人格的不断提升。所以，为破解智能机器大量使用对人际关怀、道德责任的消弭，人工智能时代的教育要向其本真回归，通过师生相互敞开自己的生命空间、满怀善意和敬畏开展教育性的真理探问[1]，来唤醒学生的生命自觉，让其体悟应该承担的责任和应该给予他者的关怀，从而在扮演自身角色的过程中不断促进自我道德的完善和生命的成长。

第三，提高人明辨是非的道德推理能力。在人工智能时代，个性化算法推荐在促使信息高效传播的同时，也将人困在了“信息茧房”中，人们不再探寻真相到底如何，进入了基于立场挑选事实的后真相时代。并且，人工智能时代是一个人人都可在虚拟空间自由发声的时代，人们往往受情绪左右，盲目地发布、转发信息，导致真相被虚假信息掩盖，不仅影响了个人的道德推理和判断力，也给整个社会的核心价值理念的传播带来不利影响。因此，要通过教育提高人的理智理性，培养人追求真理的精神和敢于为真相发声的勇气，使人不再受不良媒体的煽动和误导，做人云亦云的追随者，而是对接收信息的来源、依据、完整性等进行批判性的分析，将真相从虚假信息中剥离出来，突破后真相时代的藩篱，当遇到复杂的道德场景时，不再仅仅跟着感觉走，而是基于事实真相来确定立场，通过富有逻辑的道德推理做出正确的道德判断。

第四，强化人自觉遵守智能时代道德规范的意识。人是智能机器的研发和使用主体，伴随智能机器使用而来的风险，与其说是技术本身造成的，不如说是使用技术的人造成的。因此，开展与科技伦理相关的道德教育至关重要。一方面，既要教会普通大众各种智能机器的使用方法，更要让其牢固树立安全、道德、负责任地使用人工智能的意识，在智能时代自觉遵守道德规范。智能时代背景下信任社会的建立需要每个人的参与和努力。另一方面，要教育影响人工智能时代技术安全的“关键少数”，包括人工智能研发一线的技术人员、为人工智能研究项目出资的企业家、制定人工智能有关政策法规的政府官员等，加强对以上关键少数人员的科技伦理教育，使其始终站在人类整体利益的制高点，坚守人工智能始终为人类服务的道德原则，避免算法偏见和非法滥用带来的歧视和危险，使人工智能为增进人类福祉发挥更加积极的作用。

2.科技向善：设定人工智能的道德边界

随着人工智能功用的不断提升，工具理性遮蔽价值理性的势头已然出现了。近代以来，工具理性贬抑价值理性，致使人异化的例子不胜枚举。人工智能的发展尤其充满未知性。如果未来的人工智能突破“奇点”，产生自主意识，脱离人类的控制，那么人类社会将面临巨大灾难

。因此，人类需要为人工智能的研发和使用设定道德边界，促使价值理性回归，实现工具理性和价值理性的平衡。

一方面，完善人工智能研发的道德原则。人工智能正以指数级增长的速度发展，不断突破既有的应用场域，人们很难准确预测其未来发展的边界和潜藏的危险。人类需要为人工智能发展提供指导性的道德原则，以匡正智能机器发展引发的偏差，弥补其漏洞。20世纪中期，艾萨克•阿西莫夫(Isaac Asimov)提出"机器人三定律"(Three Laws of Robotics)。近年来，各国人工智能行业学会也相继提出了人工智能发展的道德原则。2017年1月，人工智能研究者在美国加州的阿西洛马召开了“阿西洛马会议"，提出了《阿西洛马人工智能原则》("Asilomar Artificial Intelligence Principles")。欧盟于2019年4月发布了《可信赖人工智能伦理准则》(“Ethics Guide lines for Trustworthy AI”)。2019年5月，国际科技法协会(ITechLaw)组织来自16个国家的多学科专家小组编写出版了《负责任的人工智能:全球政策框架》Responsible AI:A Global Policy Framework)。然而，这些道德原则大多都是从限制或消极预防的角度提出的，鲜有为人工智能发展提供积极的道德支持。2021年9月，中国国家新一代人工智能治理专业委员会发布了《新一代人工智能伦理规范》，列出了人工智能特定活动应遵守的18项具体的伦理要求，为从事人工智能相关活动的自然人、法人和其他相关机构等提供了伦理指引。2021年11月，联合国教科文组织发布首份人工智能伦理问题全球性协议《人工智能伦理问题建议书》，这部以国际法为依据、采取全球方法制定的准则性文书，将引导全球人工智能技术向负责任的方向发展。今后，为了使人工智能在以人为本的前提下得到快速健康发展，应整合科技、伦理、法律等领域的多学科力量，构建多方参与、协同共治的科技伦理体制，逐渐完善人工智能行业的伦理道德标准。

另一方面，恪守人工智能使用的道德界限。虽然人工智能在计算速率、学习能力等方面具有人无可比拟的优越性，但在某些方面是永远不可能代替人的。尤其在情感、精神方面，人工智能的作用是非常有限的。人工智能教师能精准地辅导学生，但不能代替教师对学生进行关怀和鼓励；智能陪伴机器人能和儿童做各种智力游戏，但不能代替父母对儿女施以亲子之爱。所以，应该从道德责任履行的角度为智能机器划定使用边界，辨明哪些工作是可以由智能机器代替人去完成的，哪些工作仍是需要人自己去做的。人不能放弃自己作为技术使用者的主体责任，否则人与人的真实的天然情感将被程式化的技术所遮蔽，人们将既没有在真实世界中很好地履行自己所承担社会角色的义务，也失去了在互相交流中获得的情感回馈，必然得不偿失。

3.算法规制：维系智能时代的道德秩序

当前，算法权力持续扩张，发达国家的教育、医疗、金融等很多行业的决策都依靠算法得出。因此，为了避免算法引发的社会风险，维护稳定的社会道德秩序，必须对算法应用的各个环节予以规制。公正、道德的制度具有稳定秩序和培育德性的功能。在智能时代，要构建公正

、完善的算法法律法规体系，使算法设计、使用、监督等各环节都有法可依，使算法公正、透明，最大限度降低潜在风险，确保人工智能切实为人类福祉服务。

第一，优化算法设计，确保源头向善。算法引发的歧视、偏见等问题很大程度上和算法本身的模型有关。算法设计者要强化科技伦理意识，将以人为本的道德准则嵌入算法设计程序中，使算法模型不断优化，杜绝算法设计者故意甚至恶意操纵算法的行为发生。算法以数据为基础。如果算法的数据来源存在问题，最终得到的将是带有偏见、歧视色彩的结果。因此，算法设计者要对数据的来源、内容、处理过程进行认真审查，保证所用数据的真实性、完整性、通用性和准确性，尽可能避免得出对特定群体有歧视的结论，给社会的公平公正带来负面影响。从事算法设计的企业应当承担算法公开和解释的义务，要提高算法的透明度，使算法服务的接受者了解算法的设计目的、运行方式、潜在风险等，防止在毫无察觉的情况下侵犯个人隐私。

第二，规范算法使用，保护公众利益。算法使用者可能直接操控算法并作用于普通大众，因此，算法使用者本身是否具有善良意志，算法运行过程是否规范与社会公众利益息息相关。目前，国家已经出台了相关法律法规对算法使用进行治理。如2021年12月，国家网信办等四部门联合发布的《互联网信息服务算法推荐管理规定》明确要求，算法推荐服务提供者应当坚持主流价值导向，积极传播正能量，不得利用算法推荐服务从事违法活动或者传播违法信息，应当采取措施防范和抵制传播不良信息［1］，为短视频、社交、电商平台等算法推荐的主要使用者划出了“红线”，从源头上减少负面低俗信息的消极影响，同时赋予用户自主选择的权利，用户可自主关闭算法推荐服务。这对于促进社会核心价值理念信息的传播、培育公众的道德共识具有重要意义。此外，算法使用者应本着负责任的态度，在数据收集时遵循准确、够用的原则，避免数据采集过度从而侵犯用户的隐私权。算法使用过程中要加强对数据的审查，剔除带有偏见的数据，同时要遵守数据的存储和使用规范，防止数据泄露威胁用户的人身和财产安全。

第三，强化算法监督，降低算法风险。政府相关部门要设立专门的算法治理机构和分类分级监管机制，对算法进行全链条动态监管，使算法相关产业负责任、可持续地发展。一旦发现违反规定的行为，就要对相关责任主体予以处罚甚至追究刑事责任。算法设计、使用的相关单位和部门应强化算法伦理责任，在内部设立算法伦理委员会，按照目前国际国内通用的算法设计、使用规范定期自主开展风险防控审查，尽早发现算法设计存在的技术漏洞和算法使用中可能引发的技术、伦理风险，确保算法始终为人服务。另外，还应组织技术专家、科技伦理学家成立第三方机构，对算法的设计、使用开展事前和事中审查，从而对技术不确定性引发的各类风险进行有效防范，对算法使用不当造成的不良影响进行及时干预。

4.和谐共生：创造人机共融的美好未来

人工智能发展带来了巨大的红利，也伴生着一系列道德风险。为了实现人工智能推动经济社会发展和保持社会道德秩序稳定的双赢局面，并促进人机协调、可持续发展，建立和谐共生的人机关系已经势在必行。

第一，要赋予人工智能道德感，构建人机共通的价值体系。早在1960年，控制论领域的创始人、传奇数学家诺伯特•维纳(Norbert Wiener)就曾说“如果我们使用一种我们无法有效干预其操作的机器来实现我们的目的……我们最好确定，机器所执行的目的就是我们真正想要的目的。”最令人担忧的问题恰恰在于人工智能的发展充满极大的不确定性，智能机器未来可能会脱离人类控制，做出违背人类社会伦理道德的事情。想要避免这种潜在危险，唯一的办法就是让智能机器拥有人类社会普遍认可的价值理念，当智能机器充分了解了人类的价值观，就不会做出危害人类的事情。为此，人类首先要思考，在智能机器全方位融入人类生产生活、精神文化世界之后，人类的道德情境会发生哪些新的变化，潜藏哪些道德风险。人类自身需针对这些新的变化主动对原有的道德规范做出调整。其次，要通过合乎道德规范的算法设计将人类社会普遍认可的道德原则嵌入智能机器的设计中，赋予其“道德感”要让智能机器进行逆向强化学习(inverse reinforcement learning，IRL)，通过直接观察人类行为并参考有关人类行为反应的书面和视频信息，通过道德算法对人类之间的利益冲突予以权衡。一旦怀有不良企图者给智能机器发送违背道德的指令，智能机器就做出应有反应，并拒绝执行。

第二，要从“关系转向”的视角重塑人和智能机器的新型关系。通常，人们总是从人类中心的视角来看待人和智能机器的关系，将其看作满足人类需求的工具。然而，人与智能机器可实现人机耦合、人机交互和人机互补已是不争的事实。人能给智能机器发送行动指令控制其行为，智能机器亦可以用其分析结果影响人的决策和行为。这使得人工智能时代的人机关系已经超越主客二元的人机关系。技术哲学家彼得-保罗•维贝克(Peter-PaulVerbeek)认为，智能人造物在某些情境下是可以表现出道德行为，起到类似主体作用的。人与智能机器构成了一种前所未有的主体与类主体关系，或者说是跨人际主体间关系。某一存在物是否具有道德地位不是取决于其主观或内在属性，而在于可观察到的客观的外在关系。并且，人工智能体是可以识别人类情绪并予以适当回应的。例如，有些老人长期接受人形护理机器人的照顾和陪伴，就会对机器人产生情感依恋。所以，人和智能机器建立伙伴关系(buddy relationship)是一种最为理想的状态。当类人机器人通过频繁的人机交互和广泛的切入点全面融入人类生活后，它们将成为人类的新伙伴。因此，人类需要从关系视域出发，以道德为指向，既要善待自己，也要善待人工智能体，唯有人机和谐共处，才能为人类赢得一个更加美好的未来。

参考文献：

［1］冯永刚、屈玲:《ChatGPT运用于教育的伦理风险及其防控》，《内蒙古社会科学)2023年第4期。

［2］曲蓉：《这一年，人工智能“生成”精彩》，《人民日报海外版》2023年12月28日。

［3］冯永刚、陈颖：《智慧教育时代教师角色的“变”与“不变”》，《中国电化教育》2021年第4期。

［4］希拉里•普特南:《理性、真理与历史》，李小兵、杨莘译，辽宁教育出版社1988年版，第248页。

［5］《普列汉诺夫哲学著作选集》第1卷，三联书店1962年版，第551页。

［6］陈彬:《历史与现实:科学技术与道德之间关系的二维考察》《理论学刊》2015年第10期。

［7］王春:《绍兴警方侦破全国首例利用AI犯罪案》，《法制日报)2017年9月26日。

［8］马克思恩格斯文集》第2卷，人民出版社2009年版，第580页。

［9］于泽元、那明明：《人工智能时代教育目的的转向》《中国电化教育》2022年第1期。

［10］贝赫鲁兹•佛罗赞:《计算机科学导论》，吕云翔等译，机械工业出版社2020年版，第149页。

［11］郭小平、秦艺轩：《解构智能传播的数据神话：算法偏见的成因与风险治理路径》，《现代传播（中国传媒大学学报）》2019年第9期。

［12］李本:《美国司法实践中的人工智能:问题与挑战》，《中国法律评论》2018年第2期。

［13］余源培等:《哲学辞典》，上海辞书出版社2009年版，第298页。

[14]H.Jonas，The Imperative of Responsibility:In Search of an Ethics for the Technological Age，Chicago:The University of Chicago Press，1984，p.90.

［15］程承坪:《论人工智能的自主性》,《上海交通大学学报(哲学社会科学版)》2022年第1期。

［16］曲蓉:《破解人工智能道德治理中的责任难题》，《中国社会科学报》2021年12月22日。

［17］斯宾诺莎:《伦理学》，贺麟译，商务印书馆1983年版，第4页。

[18]J.Savulescu，H.Maslen，"Moral Enhancement and Artificial Intelligence:Moral AI?"，J.Romportl，E.Zackova，J.Kelemen(eds.)，Beyond Artificial Intelligence.Topics in Intelligent Engineering and Informatics，vol.9，Springer，2015，p.80.

[20]黄各:《人工智能道德增强:能动资质、规范立场与应用前景》，《中国社会科学院大学学报》2022年第5期。

[21]陈万球:《技术侵入:道德自由的传统与超越》，《伦理学研究》2020年第3期。

[22]肖祥:《公民道德信任建立析论》，《中国特色社会主义研究》2019年第6期。

[23]杨先顺、莫莉:《人工智能传播的信任维度及其机制建构研究》，《学术研究》2022年第3期。

[24]唐汉卫:《人工智能时代教育将如何存在》，《教育研究》2018年第11期。

[25]曾建平、黄以胜:《信息技术问题的道德治理》，《华东师范大学学报（哲学社会科学版）》2022年第2期。

［26］王果、李建华：《人工智能时代“他-我”师生关系的建构——在教育性对话中深化责任、关怀和人格感召》，《中国教育学刊》2021年第7期。

[27]冯建军:《网络公民教育:智能时代道德教育的新要求》，《伦理学研究》2022年第3期。

[28]何怀宏:《人物、人际与人机关系——从伦理角度看人工智能》，《探索与争鸣》2018年第7期。

[29]孙伟平、李扬:《论人工智能发展的伦理原则》，《哲学分析》2022年第1期。

[30]《（新一代人工智能伦理规范）发布》，2021年9月26日，https://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html。

［31］张璁：《规范算法推荐，保障用户知情权选择权》，《人民日报》2022年1月6日

[32]王东、张振：《人工智能伦理风险的镜像、透视及其规避》,《伦理学研究》2021年第1期。

[33]N.Wiener，"Some Moral and Technical Consequences of Automation"，Science，1960，131(6)，pp.1355-1358.

[34]S.Russell，"Should We Fear Super smart Robots?"，Scientific American，2016，314(6)，pp.58-59.

[35]P.-P.Verbeek，"Materializing Morality:Design Ethics and Technological Mediation"，Science，Technology，&Human Values，2006，31(3)，pp.361-380.

[36]程广云：《从人机关系到跨人际主体间关系——人工智能的定义和策略》，《自然辩证法通讯》2019年第1期。

[37]David J.Gunkel，"Perspectives on Ethics of AI:Philosophy"，Markus D.Dubber，Frank Pasquale，and Sunit Das(eds.)，The Oxford Handbook of Ethics of AI，Oxford:Oxford University Press，2020，p.547.

[38]Misun Chu and Seoungho Ryu，"How to Embrace Artificial Intelligence?Focusing on Goffman's Theory"，International Conference E-Society，2019，pp.243-250.

[39]CathrineHasseandDorteMarieS0ndergaard(eds.)，Designing Robots，Designing Humans，London andNew York:Routledge，2020，p.2.