人工智能对个人自主性的挑战及其伦理治理

杜严勇

摘　要：自主性是保障个人幸福的关键性因素之一，历来受到伦理学、心理学等学科的普遍关注。在世界各国与国际组织发布的人工智能伦理与政策文件中，自主性原则是核心伦理原则之一。人工智能既存在提升与增强个人自主性的潜力，也可能限制并威胁个人自主性，其具体影响及其程度取决于技术设计、使用场景、使用方式与用户能力等多种因素。当前人工智能对个人自主性的负面影响主要集中在干预与依赖导致的家长主义风险、限制与操纵导致的黏性陷阱风险、约束与顺从导致的虚假自主性风险等几个方面，需要我们从明确个人自主性的让渡边界、推行积极技术设计、提升个人数字能力与素养等多个层面进行伦理治理，从而实现人工智能向善。

关键词：人工智能；伦理风险；自主性；伦理治理

中图分类号：B82　　**文献标识码：**A　　**文章编号：**1003-0751(2025)08-0105-09

人工智能的快速发展与广泛应用引发了一系列深刻的伦理问题，已经得到社会各界的广泛关注与普遍重视。不过，就人工智能引发的伦理风险而言，关于隐私保护与可信任、透明性与可解释性、公正性与安全性、权利与责任等问题已有较为丰富的研究成果，但关于个人自主性的讨论却较少。从学术史的角度看，关于自主性的研究一直是哲学领域讨论的话题之一。尽管国内外学术界对自主性的内涵仍然是众说纷纭，但对其重要意义已经达成普遍共识。因此，有必要在澄清自主性内涵的基础上，较为全面地探究人工智能对个人自主性影响的具体表现形式，提出相应的解决策略，使人工智能更好地为人类服务。

一、自主性的重要性及其内涵

自主性是影响个人幸福感的关键因素之一，也是应用伦理学的核心原则。不同时期的学者从内在论与关系论、描述性与规范性等角度对自主性的内涵进行了阐释，从另一个侧面证明了自主性的重要意义。不过，国内外发布的关于人工智能伦理原则与指南的各类文件中，对自主性的内涵界定倒是颇为相似。

1.自主性的重要意义

政治哲学、伦理学、心理学等多个学科的学者从不同角度强调了个人自主性的重要性。英国19世纪的思想家约翰·斯图亚特·密尔(John Stuart Mill)在其著作《论自由》中虽然没有使用“自主性”这一概念，但他所提到的个性与本文讨论的自主性在很多方面是重合的。密尔认为“个性自由的发展乃是幸福的首要要素之一”[1]。虽然当今社会已进入人工智能时代，但我们仍然可以认为，个人自主性是个人幸福的关键性因素之一。

心理学中的自我决定理论将自主性、能力和人际关系视为人类的基本心理需求，也是提升个人幸福感的重要因素。当个人的行为是自主的，即行为符合个人的意愿和价值观时，个人的内在动机是最强的。外部事件(如奖励或反馈)对内在动机的影

收稿日期：2025-05-26

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”(20&ZD041)。

作者简介：杜严勇，男，同济大学人文学院特聘教授、博士生导师，国家社会科学基金重大项目首席专家(上海　200092)。

响取决于它们是否支持个人自主性。支持自主性的事件(如自主选择和积极反馈)可以增强内在动机,而控制性的事件(如威胁或奖励)则会削弱内在动机[2]。心理学的实证研究表明,自主性与个人幸福感之间存在密切关联。通过对个人自主性与主观幸福感之关联的大量研究成果的综合性分析发现,个人自主性需求与主观幸福感之间存在中等程度的相关性,而且与文化背景相关性不显著[3]。也就是说,个人自主性的满足对于主观幸福感具有普遍的重要意义,这种相关性不受文化背景的影响,自主性是人类的普遍性需求。

自主性原则在伦理学领域中也受到普遍重视。有学者认为,自主准则是应用伦理学中的首要准则,指人们运用自己的理性而不是听任于异在权威或传统诱导的行为方式[4]。科技伦理通常被归为应用伦理学,因此自主性原则是科技伦理的基本原则之一。比如,自主性一直是生命伦理学的一条核心原则。有学者把"尊重"列为生命伦理学的第一条基本原则,而"自主性"则是尊重原则中的第一个具体表现。自主性是一个人按照她/他自己选择的计划决定她/他的行动方针的一种理性能力。自主的人不仅是能够思考和选择这些计划,并且是能够根据这些考虑采取行动的人[5]。有学者则把自主(允许)原则列为解决生命伦理学难题的第一原则,甚至认为它也是伦理学的第一原则[6]。

在人工智能时代,人们对自主性原则的关注有增无减。2018年,通过对979名来自各个领域的专家的调研发现,大家普遍担心人工智能在提高工作效率的同时,对人类的自主性与主体性形成威胁。通过对全球37个国家的200份涉及人工智能伦理原则与伦理指南的文献进行统计研究发现,有107份文献提及与自主性原则相关的内容,总数位于17类伦理原则中的第5位[7]。具体的实证研究表明,如果用户在与算法互动的过程中能够保持一定的自主性,如能够修改算法的预测结果(即使只能进行很小的调整),用户就会更愿意使用算法,即使他们很清楚算法并不完美[8]。也就是说,无论是当前社会中的人工智能治理政策导向,还是人工智能应用的现实关切,自主性都占据着举足轻重的地位。

2.自主性的内涵

正是因为自主性的重要意义,不同时期、不同学者从不同角度进行了风格迥异的阐释。目前国内外学术界对其具体内涵尚未形成普遍共识,使得其成为一个颇为复杂的概念。总的来看,学术界对个人自主性内涵的代表性界定大致可以分为内在论与关系论、描述性与规范性等维度。

学者们对内在论自主性关注较多。传统的自主性概念通常被理解为"个人按照自己的意愿和价值观行事",就是典型的内在论进路。从重视实现个人自主性的条件角度看,内在论自主性通常强调个人能力的重要性,特别是个人控制的能力。有学者认为,自主性一般被理解为个人根据自己的理智推理和动机欲望过自己生活的能力,而这些理智推理和动机欲望都不是操控性和歪曲性的外部力量的产物[9]。有学者注重从自我控制的角度定义自主性,认为自主性是一种自我控制的能力,这种能力允许个体根据自己的理由和价值观行动,而不受外部强制或操纵;自主性不仅仅是简单的自我决定,而且是一种能够反思自身动机和目标的能力,这种能力使得个体能够在面对外部影响时保持独立性,并根据自己的意愿做出选择[10]。为了更好地探讨人工智能对个人自主性能力的影响,有学者将自主性分为决策自主性与创造性自主性两个层级,而且这两个方面的自主性都需要得到保护。决策自主性指个人做出自主选择的能力,即不受限制或控制的能力;创造性自主性指以自主方式产生新事物的能力[11]。当然,也有学者认为,我们不能将自主性与独立做事或拒绝控制简单联系起来,因为在某些场景中个人可能乐于放弃个人控制权而接受某种依赖。

从评价或判断是否实现个人自主性的角度看,可以将自主性分为程序理论和实质理论两个方面。程序论认为,内容中立的程序为自主性提供了充要条件。例如,一个程序性理论可能会主张,只要个人基于她/他通过批判性反思所认可的实际身份或高阶欲望来选择某个目标,那么这种选择就属于自主意愿。但批评者指出,个人的身份或欲望可能是社会压迫的产物,可能本身是不合理的。实质论认为,如果人们从正确的价值观出发或按照真实有效的规范行事,他们就是自主的。实质论自主性设定了一些实质性条件,要求意愿的内容或支撑这些意愿的价值观必须满足这些条件,才能被视为自主的[12]。

尽管内在论自主性受到学者们的普遍关注,但自主性的实现与评价都是在具体的社会环境中展开的,因此自主性必然需要处理个人与外部世界的各种关系。内在论自主性是个人自主性的基础与前提,但仅仅关注内在的自主性显然是不够的。正如有学者强调的那样,"物质的、社会的和文化的环境是实现自主性的条件,而个人的自主状态也是在这

些关系中实现的”[13]。因此我们需要更多地关注关系论维度的自主性。

还有学者主张把个人自主性分为描述性自主性和规范性自主性两种。描述性自主性关注的是个体在实际决策过程中所表现出来的自主能力,如内在动机、反思性等,或者说个体是否具备某些条件或特征以实现其自主行为。描述性自主性又可以分为内部主义和外部主义两种进路。内部主义强调自主性与个体的内在动机、欲望、长期计划或自我认同密切相关,只有当个体行为是基于个体自身认可的动机或欲望时,其行为才能被视为是自主的。外部主义则强调自主性需要通过某种外部标准或理由来验证,只有当个体行为能够满足充分的理由,并且这些理由与个体的长期利益或目标一致时,其行为才能被视为是自主的。规范性自主性指个体作为自主个体的权利,个人应被视为目的而非手段,其利益和偏好不应被任意侵犯或干涉。规范性自主性要求,个体的自主性应受到尊重,其决策和偏好不应被随意干预;个体有权根据自己的意愿和偏好做出选择,而不应受到外部的强制或操纵。

描述性自主性是规范性自主性的基础,个体需要具备一定的描述性自主性,才能被认为拥有规范性自主性;在某些情况下,个体的描述性自主性可能受到损害(比如被操纵或误导),但他们仍然拥有规范性自主性,也就是应该受到尊重[14]。从前述内在论与关系论的角度看,我们可以把描述性自主性与规范性自主性看作是内在论与关系论的精简版本,描述性自主性(特别是内部主义)属于内在论自主性,而规范性自主性属于关系论自主性。

3.人工智能伦理原则与政策对自主性的重视及其界定

如前所述,大多数人工智能伦理原则、伦理规范与指南等文件中提到了要保护、尊重、实现个人自主性,部分文件中还提出了相应的举措。以下简要列举几个代表性文件中提及的相关内容加以说明。

2018 年,蒙特利尔大学牵头发布的《蒙特利尔负责任开发人工智能宣言》中提出了 10 条人工智能原则,其中第 2 条是“尊重自主性原则”,强调人工智能的开发和使用必须建立在尊重个人自主性的基础上,以加强人们对生活和周边环境的控制为目标,帮助个人实现道德目标和有意义的生活,同时不能通过各种手段将特定的生活方式强加于个人。

2019 年,欧盟高级别专家组发布的《可信任人工智能伦理指南》中提出了 4 条伦理原则,其中第 1 条就是“尊重人类自主性”,强调人们在与人工智能系统互动的过程中必须保持完整且有效的自主决定权,人工智能系统不应该在无正当理由的情况下使人类处于从属地位、施加胁迫、进行欺骗与操控、制约与驱使人类。2020 年,欧盟发布《人工智能白皮书:通往卓越与信任的欧洲进路》,提出了针对人工智能的监管框架,其中强调了对高风险人工智能进行“人类监控”的重要性:“人类的监控有助于确保人工智能系统不会破坏人类的自主性或造成其他不利影响。只有通过确保人类对高风险人工智能应用程序的适当参与,才能实现可靠、合乎道德和以人为中心的人工智能的目标。”

2021 年,联合国教科文组织发布《人工智能伦理问题建议书》,提出的“人类监督与抉择”原则涉及人的自主性问题,强调在某些情况下(比如出于效率方面的考虑),人类可能选择依赖人工智能系统,但是否在有限情形下让渡控制权,依然由人类来决定。一般来说,生死攸关的抉择不应该交给人工智能系统。

2021 年,中国国家新一代人工智能治理专业委员会发布的《新一代人工智能伦理规范》提出了 5 条基本伦理规范,其中第 4 条是:“确保可控可信。保障人类拥有充分自主决策权,有权选择是否接受人工智能提供的服务,有权随时退出与人工智能的交互,有权随时中止人工智能系统的运行,确保人工智能始终处于人类控制之下。”虽然此条伦理规范被称为“确保可控可信”,但实质上就是“自主性规范”。2023 年,中国国家人工智能标准化总体组、全国信标委人工智能分委会发布《人工智能伦理治理标准化指南》,在提出的第 1 条伦理原则“以人为本”中,涉及人的自主性内涵。在关于“自主自由”可能的应用场景中,提到“医疗保健、教育和司法”3 个方面,而且基本上都强调人类的选择权与控制权。

从以上几份代表性文件的简述可见,国内外发布的相关文件中,对人类自主性的论述几乎一致性地强调人类应当掌控人工智能系统,应该充分保障个人的知情权、决策权、自主选择权。根据前文对自主性的内涵介绍可见,人工智能政策文件中的自主性大都属于关系论自主性与规范性自主性。

二、人工智能对个人自主性的挑战

前人工智能时代的个人自主性,主要是建立在人与自然、人与动物相区别的基础上,从人与人、人

与社会之间相互影响的角度展开的。科学技术的发展对个人自主性的影响固然有消极的一面,但基本上以积极的为主,因此人机关系的探讨较少与自主性相关联。不过,随着人工智能的快速发展与广泛应用,而且由于其智能程度不断提高,自主程度也日益增强,对个人自主性形成了威胁与挑战,因此人工智能的自主性与个人自主性的矛盾逐渐凸显并加剧,已经成为处理个人与人工智能关系的核心议题。毋庸置疑的是,人工智能可以通过增强个人能力来提高日常生活与工作中的自主性,使人们更方便、快捷、高效地实现自己的任务与目标。同时,人工智能引发的自主性挑战也日益深刻,可能产生不尊重用户的自主性、引导用户追求更少或更低价值的目标、引发更糟糕的自主性状况等风险,影响内在论与关系论、描述性与规范性等各个层面的个人自主性。以下结合具体的人工智能应用场景进行分析探讨。

1.干预与依赖:家长主义风险

在医疗领域中,医生或其他医疗专业人员通常比患者更了解患者的健康状况,他们基于对患者最有利的立场,代替患者做出决策,这种行为通常被称为“家长主义”,而家长主义被认为是对患者自主性的限制。医学伦理史上的一个重要变化就是从家长主义转向尊重人们的自主性。比如,美国医学会的《医学伦理守则》,早期要求患者无条件接受医生的处方,而现在则明确要求医生尊重患者的抉择,强调有抉择能力的患者可以接受,也可以拒绝任何医疗干预[15]。

目前,人工智能在医疗保健领域得到日益广泛的应用,人工智能系统在没有或只有极少数患者参与的情况下,独立做出决策,由此产生了“人工智能家长主义”现象。家长主义也有软硬与强弱之分。软硬家长主义主要是针对干预的强度进行区分的。软性家长主义指干预者暂时限制或影响个体的选择或行动,目的是确定该选择或行动是否真正自主。一旦确认个体的选择是自主的,干预者会允许个体按照自己的意愿行事,其核心在于尊重个体的自主性,仅在必要时进行最小限度的干预。硬性家长主义指干预者直接限制或否决个体的自主选择,认为自己比个体更清楚什么对个体有利。强弱家长主义的差异在于干预的是手段还是目标。弱家长主义指干预者对个体实现自主选择目标的手段进行干预,而不是直接干预目标本身。干预者认为个体选择的手段可能无法实现其自主选择的目标,因此提供帮助或限制手段的选择。强家长主义指干预者直接干预个体自主选择的目标本身。这种干预通常涉及对个体生活方式、价值观或人生目标的根本性改变,认为个体选择的目标是错误的或有害的[16]。

从上述对家长主义的分类可见,除极个别特殊情况外,硬性家长主义、强家长主义是需要明确反对的。需要强调的是,传统家长主义的主体是人或社会组织,比较容易识别,人工智能家长主义则更具有隐蔽性。虽然软性与弱家长主义对不少人来说是可以接受的,但我们也应该对其保持必要的警惕。从对个人信息与行为干预的角度看,以下几种软性家长主义与弱家长主义需要予以重点关注:一是削弱个人的知情同意权。人工智能系统可能出于系统设置,有选择性而不是毫无保留地向个人提供信息。二是代替个人进行决策。由于技术因素或个人原因,个人无法参与决策过程,人工智能代替个人进行决策,限制个人的选择权,忽视个人的意愿与偏好。三是在合理目标指引下产生的不恰当的行为控制与引导。比如,在医疗保健领域中,人工智能可能出于维护患者健康的目标,忽视患者自主意愿,要求其采取服药、运动或饮食控制等行为。

传统医疗领域的家长主义通常对干预强调较多,对依赖关注较少。在很多人看来,人工智能比人类专家更客观、更理性,因此他们更愿意接受人工智能的建议而不是人类专家的建议,甚至会产生对人工智能的依赖现象。我们认为,与家长干预孩子的情况相联系的现象是,孩子对家长也会产生某种程度的依赖心理与行为,而家长也乐于接受这种依赖。依赖主要是从孩子的角度出发的,但从孩子个人成长的角度看,家长纵容孩子的过分依赖也是不恰当的。因此,纵容个人过度依赖人工智能也应该被看作是另一种家长主义的表现。有学者认为,如果个人过度依赖人工智能的建议,那么她/他可能逐渐失去独立做出决策的能力,在很多方面都越来越依赖人工智能,产生所谓的“婴儿化”现象[17]。研究表明,生成式人工智能对创造性自主性会产生负面影响。如果用户过度依赖生成式AI来完成创意任务,可能会导致自身创造性能力的退化。例如,长期依赖人工智能生成写作内容的用户可能会逐渐失去独立创作的能力[11]。

2.限制与操纵:黏性陷阱风险

人工智能推荐系统通过算法对用户进行分类和建模,限制用户的选择范围,从而影响个人自主性。通过把握用户的世界观与动机,推荐系统作为“黏性陷阱”发挥作用,它们试图将用户“黏”在某些特

定的解决方案上[18]。由此可能对用户的自主性产生以下几个方面的影响。

第一,限制用户的选择范围。推荐系统通过分析用户的点击浏览行为,构建较为精准的用户个人画像,根据用户的兴趣与偏好,选择更符合用户预期的信息,同时过滤掉人工智能系统认为无关或用户不感兴趣的内容,由此会在很大程度上限制用户接触到的信息范围,从而减少用户的选择多样性。虽然这种做法可能会提高推荐的相关性与准确性,但也可能让用户陷入“信息茧房”,使其逐渐失去接触不同观点与信息的机会,导致用户对世界的理解偏于片面化与单一化。

第二,操纵用户偏好。通过算法干预,推荐系统可能会逐渐改变用户的偏好,使用户更倾向于接受系统推荐的内容。这种偏好操纵可能会让用户在不知不觉中接受某些观点或产品,而没有意识到自己的偏好已经被改变。在这种情况下,用户的自主性并没有被否定,而是被有效地利用和干预了。研究表明,推荐系统会产生锚定效应,影响用户的偏好。锚定效应是指在决策或判断过程中,人们会不自觉地依赖于某个初始值(即“锚点”),并在此基础上进行调整以形成最终判断,从而导致判断结果偏向于该初始值的现象。在推荐系统中,消费者对产品的偏好评分会受到推荐系统提供的预测评分的影响,即使这种评分并非完全客观,消费者仍会将其作为参考依据,进而调整自己的评分,表现出显著的锚定效应[19]。锚定现象可能会导致个人逐渐偏离自己认同的价值观,甚至产生个人偏好的变化,形成适应性偏好,由此破坏个人自主性。

第三,重塑个人身份。推荐系统通过用户分类和标签化处理,将用户归类到特定的社会类别中,并根据这些类别提供推荐,特别是通过提供令用户愉悦的信息,改变用户的行为习惯,使其更倾向于某些特定行为,行为的调整反过来会改变用户对自己身份的认知。个人自主性强调个人的行为应当反映其真实动机与价值观,而个人身份正是这些动机与价值观的体现[14]。因此,个人身份重塑的结果可能与用户的自我认知不一致,使用户对自己的身份与偏好产生误解,影响其自主决策能力,由此可能对用户的长期发展产生负面影响。

按照前述关于自主性的哲学观点,作为“黏性陷阱”的推荐系统侵犯了用户的自主性,它不仅限制了用户获取的信息类别与质量,还深刻地塑造了用户的个人行为及其对自己与社会的认知。从个人自主性的不同层面的角度看,推荐系统既影响了内在论的个人自主性,也影响了关系层面的个人自主性,它限制了个人接触社会的广度和深度,阻碍个体接触更为广泛的社会现实。

3.约束与顺从:虚假自主性风险

2018 年 4 月,欧盟发布了一项关于在数字化世界中推动健康和护理数字化转型的政策文件,其主要目标在于通过数字化手段提升健康和护理系统的效率、可及性和可持续性,同时赋予公民更多的控制权,鼓励公民参与健康管理,实现个性化的健康与护理。我们很容易发现,近年来,“赋权”在许多国家政策文件中经常出现,但是,赋权并不一定就会提高公民的自主性,反而可能会起到负面影响。现有政策文件中大多并未详细说明数字健康工具如何真正赋权于公民或患者,相反,这些数字技术可能迫使个人通过“数字化医学凝视”对自己生活的各个方面进行自我监控,而非关注于如何利用数字健康工具产生的数据改善自己的医疗服务。“数字化医学凝视”指个体从医生的视角出发,利用数字健康工具提供的信息进行自我评估的实践活动。因此,个体被鼓励对照健康基准评估自己的表现,但这些基准的确立依据、适用性及适用范围却未被明确告知。而且,数字健康工具提供的建议往往倡导“顺从”,而非“自主”[20]。

具体而言,“赋权”对用户自主性的影响主要体现在以下几个方面。

第一,规范性约束取代自主选择。数字健康工具通过预设健康标准(如每日步数、卡路里限制)将“健康”定义为可量化的目标,用户被鼓励通过“数字化医学凝视”不断自我监控。这种标准化框架将健康窄化为对数据的服从,用户的选择自由被限制在工具设定的路径内,自主决策空间被压缩。比如,若数字健康工具要求用户每日完成万步目标,即便个体因身体条件或环境限制无法达标,仍会被系统标记为“不健康”,迫使用户产生焦虑或自我责备,而非根据实际需求调整行为。

第二,责任转移与结构性忽视。赋权叙事将“患者”重构为主动的“用户”,暗示健康结果完全由个人选择决定。当个人无法达到健康标准时,责任往往被归咎于个人不够努力,而不是个人特殊情况或环境因素。这种责任转移可能使个人感到内疚,从而削弱了他们的自主性。另外,赋权模式过于关注个人的决策能力,而在很大程度上忽视了其行动能力。即使用户能够做出合理的决策,他们也可能

因为收入、教育水平、客观环境等因素无法将决策转化为实际行动。

第三,算法操控与自主性幻觉。数字健康工具通过算法“助推”引导用户行为,比如,通过推送通知、振动提醒等手段,促使个人采取某些行为(如运动或饮食控制)。这种引导可能使个人在没有充分反思的情况下做出决策,因此削弱其自主性。这些设计虽以“透明”自辩,实则通过心理学机制操控选择,用户通常难以察觉其决策如何被技术干预。用户往往误以为行为源于自由意志,其实是算法对“数字自我”的持续调控,这种隐蔽的控制削弱了用户对自身健康管理的真实掌控[21]。

另外,数字健康工具可能在用户不知情(或不完全知情)的情况下收集个人数据,这种被动的数据收集导致用户与商家之间的信息不对称,用户不知道哪些或哪些类型的数据会被用作何种用途。通常情况下,个人可以决定上传哪些方面的信息,在个人信息的展示场合、模式与数量等方面拥有自主权,但被动的数据收集限制了用户的自主性,削弱了用户对个人信息的控制权。

三、多主体共同推进对自主性风险的伦理治理

通过上文的论述可见,当前个人自主性的各个维度已经不可避免地受到人工智能的影响。因此,人们智能时代的个人自主性是用户与人工智能相互作用、协同进化、共同竞争的结果。就目前人工智能的技术发展水平以及前文所述的主要风险而言,需要哲学家、科研人员与社会公众等所有利益相关者共同努力,重点从以下几个层面进行伦理治理,以规避或降低人工智能对个人自主性的负面影响。

1.工具与代理:明确委托与授权模式,澄清个人自主性的让渡边界

人工智能拥有日益增强的自主性,人类也确实越来越多地将某些决策交给人工智能。因此,人工智能协助人类进行判断与决策,不能简单地都视其为干预与控制。从前述关于描述性自主性与规范性自主性的界定来看,我们需要明确人机分工机制,充分保障人类的基本权利。人工智能是帮助人类实现某些目标的工具,但在实现目标的具体过程中,一些决策过程及其细节,人类不进行干预,也可以说是人类授权人工智能进行决策。或者说人工智能是人类的代理,人类把某些决策的权利移交给了人工智能[22]。因此,个人与人工智能之间形成了权利让渡(或授权)与权利受托的默认关系。

就个人实现具体的目标而言,将具体执行的过程与细节交由人工智能进行决策,并不会妨碍或削弱个人自主性,反倒是提高个人自主性的重要手段。只要授权在个人的控制之下,同时符合个人的利益与偏好,就不会对个人自主性形成威胁。也就是说,人类对人工智能进行委托与授权已成为客观现实,我们面对的关键性问题是澄清个人自主性让渡的边界。卢恰诺·弗洛里迪(Luciano Floridi)认为,我们需要明确人类的“元自主性”,即人类应当保留决定采取哪些抉择的权利,在必要时可以进行自由选择;除一些特定情况外,人类一般不让渡选择权,而且对任何让渡的选择权都应当保持收回的可能性[23]。但是,弗洛里迪并没有论证具体应当保留的权利内涵。一般而言,人们最关心的基本权利至少包括生命权、隐私权、知情同意权、自主选择权和对人工智能的控制权。人们普遍希望,当个人感觉自己的生命权、隐私权等基本权利受到威胁时,能够对人工智能进行有效的控制。更重要的是,在不同的应用场景中,不同的权利与价值排序及其实现模式存在巨大差异,需要具体问题具体分析,不能一概而论。

如何在不同的应用场景中明确人类元自主性的具体内涵,必须要加强伦理学家与科学家的合作,但关键是如何合作的问题。很多学者反复强调,人工智能伦理问题的解决需要科学家与哲学家密切合作。没有科学家的支持,哲学家的理论会过于抽象而无法实施;反过来,没有哲学家的配合,科学家可能会对伦理理论进行错误的阐释[24]。可以将苏格拉底对话作为一种方法论工具,促使科学家与哲学家在反复对话的过程中,对关键术语的内涵进行澄清,揭示其深层含义,并在一定程度上形成共识[25]。通过科学家与哲学家的有效合作,可以明确不同应用场景中的价值排序,澄清维护个人自主性的有效途径,确定人工智能作为工具与代理的权利范围,进而有效化解人工智能家长主义等风险。

2.设计与体验:推行积极设计与积极计算,提升用户积极体验

近些年,在技术设计领域,科技工作者一直努力提升用户使用技术产品的积极体验,进而提高用户的幸福感,从积极技术、积极设计、积极计算等概念的提出及广泛流行即可看出这一趋势。由于自主性在用户幸福感中的基础性地位,因此重视用户自主性是技术设计领域关注的一个核心问题。从人机交

互的伦理设计角度看,价值敏感设计的开创者之一巴蒂亚·弗里德曼(Batya Friedman),早在1996年的论文中就强调要重视两个价值观:用户的自主性与避免偏见。他认为,可以从系统能力(系统拥有实现用户目的所需的各种功能)、系统的复杂性(过多的功能导致系统复杂性增加,可能降低系统的可用性,需要在功能与复杂性之间寻找平衡)、系统的错误表征(系统提供的信息不准确,或产生误导,会损害用户的自主性)、系统的灵活性(系统设置可以根据用户偏好和目标的变化而进行调整)等四个方面努力,帮助用户更好地控制技术,从而提升用户的自主性[26]。

弗里德曼强调"通过完善技术设计的理念与方法,来增强用户的自主性",其主张在科技工作者中得到普遍重视。"积极计算"的提出者和倡导者拉斐尔·卡里罗(Rafael Calvo)和多利安·皮特斯(Dorian Peters)强调:"努力尊重用户的自主性,并为此设计,对积极计算至关重要。"[27]卡里罗和皮特斯赞成弗里德曼的主张,认为我们需要在技术层面和用户体验层面都认识到自主性的重要性。设计者不能从自身的角度识别用户的需求,而应该努力让用户成为技术设计的参与者和创造者。

我们应当通过重视技术设计,实现积极人工智能的目标,确保人工智能的发展能够真正促进包括自主性在内的人类福祉[28]。从技术设计促进个人自主性的角度看,科技工作者在优化人工智能系统设计时需要重点关注三个方面的问题:一是合理控制用户参与程度。人工智能过度的推荐显然会削弱用户的自主性,但很少推荐又无法满足对社交互动与信息资讯的需求。二是平衡用户的新颖性与个性化需求。如果过度推荐新颖性的内容,可能会让用户感觉自己被故意"引导",而只推荐用户之前关注的内容,会陷入黏性陷阱风险。三是权衡长期影响与短期影响的冲突问题。人工智能往往关注即时的用户反馈,但对短期目标的关注可能与用户的自主性相冲突。

3.合作与陪伴:精细化分析人工智能的影响,构建语境化的自主性

有学者主张从六个细致的技术体验层级入手,精细化地分析技术对用户心理需求产生的影响。采用层:用户使用技术前的经验,以及促使用户使用该技术的动机;交互层:用户与技术产品的交互体验对心理需求的影响;任务层:技术促使用户发生的各种活动;行为层:用户活动服务于某种整体行为及其影响;生活层:技术对用户整体生活的影响;社会层:技术对全体社会成员的影响[29]。当然,这些层级的区分是相对的,相互之间存在交叉与关联。显然,这六个层级的分析都与用户的自主性存在密切关联,已经受到人工智能科技工作者的关注和采纳。

在精细化分析技术对自主性影响的基础上,我们需要建构一种语境化的自主性内涵,因为自主性的内容都是高度依赖于具体的社会文化或使用环境的。我们生活的各个方面既有积极主动的行动,也有被动接受的情况。受动性不应该简单地被视为对自主性的削弱,而是要区分具体的情境。比如,助老机器人在紧急情况下出于对老人安全的考虑,采取必要的干预措施是可以接受的。但是,除特殊情况之外,当助老机器人的干预与老人的自主选择之间发生矛盾时,应当充分尊重老人意见,根据老人的个人意愿来确定助老机器人的行为模式。因此,我们需要合理确定人工智能的角色与功能定位,协助用户维护个人自主性。在一些应用场景中,人工智能应当成为个人的"意志辅助工具",帮助用户实现其自主目标[30]。比如,假设用户在午餐时面临多种选择,人工智能建议选择健康沙拉而非用户更喜欢的炸鸡或薯条。尽管人工智能的建议可能使用户在短期内感到受限,但从长远来看,这有助于用户实现健康饮食的自主目标。为了避免人工智能威胁用户的自主性,作为"意志辅助工具"的人工智能应当遵循这样的原则:以用户自主设定的目标为出发点,给用户较为详细地阐述理由,同时提供多种选择,并尊重用户的最终选择。这样不但不会削弱用户的自主性,从长远来看,反而有利于显著提升用户的能力与自主性。

如前所述,数字健康工具的本意是提升用户的自主性,但却可能产生相反的效果,那么,如何应对赋权反而产生阻碍个人的自主性问题?其关键思路在于对数字健康工具进行更为准确合理的定位。有学者主张将数字健康工具从"赋权工具"重新定位为"数字陪伴",将其作为个人与医生之间的中介,根据个人不同的生命阶段与具体身体特征等个人情境动态调整医生的建议,帮助用户与医生在动态协调中实现用户的健康管理。比如,数字健康工具应该根据用户的病情阶段调整不同的干预强度,以囊性纤维化患者为例,在患者的青少年时期提供指导,成年时期转为提醒功能,这些都不是强制性的"赋权"[21]。"数字陪伴"的定位有利于医生和用户共享信息,减少信息不对称现象,既有助于用户自主判