

伦理约束推动AI技术向善而行

话题主持:本报记者 张炯强 王蔚

AI技术发展突飞猛进,当曾经的幻想越来越接近现实的时候,人们不得不产生忧虑:如果我们创造出来的AI是个“坏人”,且不受人类的约束,该怎么办?

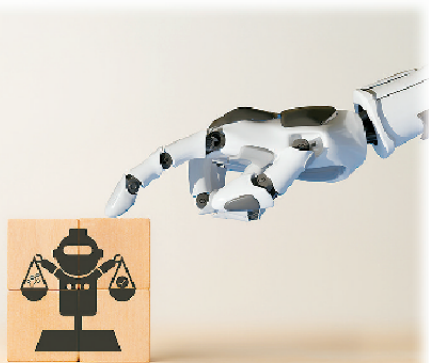
有意思的是,几十年间,好莱坞出品的关于AI的影片,竟然绝大多数是负面的。上世纪70年代,一部《未来世界》风靡全球,机器人成了与人类长得一模一样的物种,疯狂作恶;2001年,大导演斯皮尔伯格的影片《AI》,讲述了一个机器男孩与人类不可相融的感情悲剧;而在好莱坞系列电影《终结者》

中,高度智能电脑“天网”发动核战争,几乎灭绝人类。可以说,早在AI诞生之初,人们就已经为它设定了禁区,推动科技向善。而在当今,科学家们已经尝试让AI掌握自己思考能力之时,各国政府均设立了相关法规规范科技伦理加以约束。

搞科研还要伦理制约?答案是肯定的。人工智能、基因编辑、辅助生殖技术……科研领域的创新五花八门。2018年,南方科技大学的贺建奎利用基因编辑技术“造”出了天生免疫艾滋

病的女婴,舆论哗然:科学家如果可以随意“制造”自己设定的人,岂不天下大乱?贺建奎因此获刑三年。

在人类历史上,科研和发明往往是把双刃剑。未来AI的发展,如果对人类的生活方式、价值观、文明,甚至对人类存续本身产生影响、冲击或威胁,必须接受伦理制约、法律制约。新兴前沿技术发展迅速,不仅要关注技术安全性方面的风险,也要关注到人们基于道德理念分歧带来的问题。我们不需要人工智能掌控的世界,更要拒绝“天网”。



人工智能时代的科技伦理嵌入、示范与拓展

同济大学法学院助理教授、上海市人工智能社会治理协同创新中心研究员 朱悦

科技伦理嵌入开展人工智能科技活动的组织架构中

《办法》第九条规定:“从事人工智能科技活动的高等学校、科研机构、医疗卫生机构、企业等是本单位人工智能科技伦理管理服务的责任主体。有条件的单位应设立人工智能科技伦理委员会。”人工智能科技伦理委员会(以下称委员会)包括人工智能技术、应用、伦理、法律等相关背景的专家,其主要职责包括指导科技人员开展科技伦理风险评估、按要求跟踪监督相关科技活动全过程、组织开展科技伦理知识培训,等等。循此,开展相关科技活动的组织可以通过设立委员会并发挥其指导、监督、培训职责,在组织架构和活动流程中内化科技伦理。

科技伦理嵌入人工智能系统生命周期的各个环节中

《办法》第十五条规定:开展人工智能伦理审查重点关注公平公正、可控可信、透明可解释、责任可追溯和人员资质等五个方面。其中,前四方面涵盖数据的完整性和多样性、算法模型和系统设计的合理性、系统设计的鲁棒性和系统运行的持续监测,等等,并且要求“采取日志管理等措施清楚记录数据、算法、模型、系统各个环节的相关信息,保障全链路可追溯和管理,出现问题时可精准定位具体环节并确定责任人”。换言之,从开发研发时采集数据、设计系统,再到完成之后投入应用、持续运行,人工智能系统的各个环节都要符合科技伦理规范。

科技伦理在组织和技术层面深刻嵌入人工智能,体现了人工智能规范体系建设的中国立场

从全球角度看,如何安置科技伦理在人工智能规范体系中的位置,如何处理科技伦理和法律规范的关系,长期以来没有公认的答案。2024年,欧盟《人工智能法》正式通过。其序言部分第25条明确指出:人工智能的各项伦理原则一方面有助于设计融贯、可信赖、人类中心的人工智能系统,另一方面不能阻碍《人工智能法》

和其他欧盟法律的实施。换言之,法律优位于伦理,不在人工智能专门立法中设置伦理审查或类似制度。《办法》体现了另一种思路,更重视科技伦理和法律规范形成合力、双向衔接。未来,随着人工智能时代科技伦理更加重要,《办法》可能为人工智能全球治理提供示范。

结合产业和技术动向,可以探索科技伦理规范向个人用户的拓展

当前,高等学校、科研机构、医疗卫生机构、企业等单位是科技伦理规范责任主体。随着人工智能日益融入社会生活,个体用户使用人工智能的场景和频率增加。对其自身和其他用户造成损害的可能性概率和严重程度,同样相应增加。如果个体用户因使用人工智能成瘾而遭受严重损害,此时很可能需要伦理规范的介入;又如个体用户使用人工智能“复活”追思逝者,可能导致更复杂的伦理问题,等等。在人工智能时代,个体更有必要具备一定的科技伦理素养。可从素养培育入手,在已有公认伦理判断的场景中以柔性方式逐步建立规则。

结合产业和技术动向,可以探索科技伦理主体向人工智能本身的拓展

当前,人工智能系统只是伦理审查的对象之一。如《办法》第二十一条和附件一规定:对于面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统,在开展科技伦理初步审查的基础上,还需要由相关专家开展复核,确保其在公平公正、可控可信等各个方面符合要求。2024年来,国际上开始出现关于人工智能福祉的研究。换言之,承认人工智能伦理主体地位,探索界定和尊重其福祉。少数前沿人工智能企业开始在发布人工智能前实施福祉评估,甚至允许人工智能主动中止和用户的对话。当然,人工智能福祉的研究,一方面才刚刚起步,另一方面如果取得结论,在政策上、产业上、学理上可能产生很大的影响。总之,值得深入研究,目前还不成熟,有待更多的探讨。

人工智能的发展越来越迅速,引起的伦理问题也越来越受到人们的重视。从道德主体的角度来看,智能机器人无法成为真正的主体,因为它不可能拥有心智状态,而拥有心智状态是成为道德主体的必要条件。从道德接受者的角度来看,只有当智能机器人能主动地面向人类,或者某人与智能机器人发生过有意义的关联时,智能机器人才能获得道德接受者的地位。而且,这样的智能机器人必须具备对话能力,会使用某种语言。

智能机器人是否具有心智状态呢?在我看来,虽然心智状态和智能机器人一样,都依附于物理性的存在,但是,二者之间仍然存在着巨大的差异。

从本体论的角度来看,心智状态与物理存在之间有着本质性的区别。首先,心智状态的一个基本特点是具有意向性,即总是关于对象的。通常情况下,我们认为,物理性的存在并不具备意向能力。智能机器人也是如此,它无法具备意向能力。智能机器人大体可以分为两个部分,即大脑与躯体。其大脑部分的核心是算法程序,而躯体部分则是由一系列物理性存在构成。物理性存在不具备意向能力,因此,我们需要讨论的核心是算法本身是否具备意向能力。目前学界流行的算法大概有五大类,即规则和决策树、神经网络、遗传算法、概率推理和类比推理。这些算法本身虽然运行模式各异,但仍然遵从“输入—输出”模式。也就是说,只有当有信息输入时,智能机器人才会有输出。而且,这种信息正常情况下应该是真实的信息,但这明显与我们所说的意向能力不同。

从认识论的角度来看,心智状态通过理解总是体现出语义特征,但智能机器人无法达到真正的理解。塞尔的“中文屋思想试验”明确地表达了这一论点:塞尔不懂中文,但是通过一本中英文对照手册就可以将中文输入,再按照一定的规则输出中文,在外人看来,塞尔似乎懂中文,因为他的中文输出和懂中文的人的输出类似,但根据前提假设,塞尔本人不懂中文。塞尔认为,这一过程其实就是计算机的运行过程。因此,我们不能认为计算机具备理解的能力。

有些学者并不同意塞尔的论证。他们认为,目前的智能机器人并不能等同于计算机。在一定程度上,这一反驳是合理的,但是,这并不能反驳我们的最终结论,即智能机器人不具备语义理解能力。在塞尔的“中文屋思想试验”中,计算机程序被视为符号的输入与输出,而目前的智能机器人比计算机程序更复杂,其根本特征在于,它不但有符号的运行,还能与外界发生因果联系。问题在于,这两者都不能体现出语义特征。符号的输入与输出只有句法特征,没有语义特征,而因果关系同样无法体现出理解的语义特征。如果认为因果关系具备语义表征能力,那么,自然界中的很多事物都具有语义表征能力,比如动物甚至植物,因为这些对象本身受因果关系的限制,如树的年轮。

从价值论的角度来看,心智状态总是自由的状态,不受特定目的的限制。比如,它可以思考数学系统中的命题,也可以表征经验对象。但是,智能机器人不能如此,它总是受限于它从属的系统。以智能无人驾驶汽车为例,虽然可以在其系统内实现自由选择,比如从某地到某地的路线规划,但是它不能超出这一系统,去自由地选择完成别的事情,比如拿一个盒子、写一论文等。因为智能机器人总是为了某一个特定的目的被设计出来的,总是受到它的算法程序的限制。在适用于所有一切事项的终极算法被创造出来以前,智能机器人永远不可能突破它所属的系统。因此,智能机器人不可能实现真正的自由。

智能机器人该不该拥有「心智」

上海大学哲学系副教授 戴益斌

