doi:10.13582/j.cnki.1672-7835.2023.03.016

ChatGPT 技术中的人工智能伦理风险及其科学祛魅

陈元,黄秋生
(南华大学 马克思主义学院,湖南 衡阳 421001)

摘　要:作为一种生成式预训练变压器,ChatGPT 技术在推动人类社会进入通用人工智能时代的同时,也会“堆积”一些形而上学问题。其在技术演进中,呈现的偏差性、不可靠性、鲁棒性、有毒性等伦理特性,诱发了道德判定障碍、社会偏见刻板化、用户数据去隐私化、科学技术异化等人工智能伦理风险。只有凸显人文价值关怀,重塑伦理主体责任,强化伦理政策导向,创新伦理运行机制,才能有效纾解 ChatGPT 技术带来的伦理风险,使人工智能系统成功融入人类社会。

关键词:ChatGPT;“隐形的伦理代理人”;人工智能伦理风险;科学祛魅
中图分类号:B82　　**文献标志码**:A　　**文章编号**:1672-7835(2023)03-0135-08

近年来,随着人工智能算法与自然语言处理技术的发展,生成式语言模型推动人工智能技术取得了突破性进展。尤其是自 2022 年 11 月 OpenAI 人工智能研究实验室推出 ChatGPT 技术以来,人工智能系统逐渐在图像生成、信息提取、人机对话、情感分析、科学研究等维度不断满足人们用自然语言与计算机交流的愿望。然而,值得关注的是,ChatGPT 技术给人类社会带来巨大发展红利的同时,其存在的偏差性、不可靠性、鲁棒性、有毒性等隐形因素也引发了重大伦理关切。这些隐形道德因素在不同程度上诱发了道德判定障碍、社会偏见刻板化、用户数据去隐私化、科学技术异化等伦理风险。全面探索 ChatGPT 技术在伦理方面存在的“高后果风险”,不仅有助于缓解人类对于人工智能快速发展引发的道德焦虑,而且能够有效回应一个给定的道德框架能否在人工智能技术中实现的问题。因此,研究 ChatGPT 技术存在的人工智能伦理风险,有着迫切的需要。

一　ChatGPT 技术的缘起与发展

ChatGPT 技术是由人工智能研讨实验室 OpenAI 推出的一款生成式预训练变压器(Generative Pre-trained Transformer)。作为 InstructGPT 的兄弟模型,该语言模型的设计和训练得益于复杂的 Transformer 神经网络架构,巨大的语料库及人类反馈强化学习(RLHF)方法。依靠强大的算法、数据、算力,ChatGPT 技术不仅能够解决人工智能遇到的“计算机视觉、自然语言理解,以及处理真实世界中的意外情况”①等关键问题,而且能借助语言生成、上下文学习、常识和逻辑推理等能力捕捉人类的长期依赖性,灵活应对真实世界中的对话。

在对自然语言模型的创造性探索中,ChatGPT 技术呈现出一个迅速发展的趋势,其研发公司主要是谷歌与 OpenAI 两大机构。2022 年 11 月,在 OpenAI 实验室正式推出 ChatGPT 技术以前,谷歌早在 2017 年就推出了 Transformer 神经网络架构,这为后续 GPT 的迭代奠定了基础架构。2018 年,谷歌接着推出了对标 GPT 的编码器

收稿日期:2023-03-22
基金项目:湖南省重点马院重大项目(2022ZDAM04);湖南省社会科学成果评审委员会一般项目(XSP22YBZ087)
作者简介:陈元(1992—),女,湖南岳阳人,博士,讲师,主要从事马克思主义哲学研究。

①V. Roman.Yampolskiy. *Turing Test as a Defining Feature of AI-Completeness*. In Artificial Intelligence, Evolutionary Computation and Metaheuristics (AIECM)-In the Footsteps of Alan Turing. Xin-She Yang(Ed.), *Springer*, London. 2013,pp.3-17.

“BERT”,该模型拥有 3.5 亿参数。同年 6 月,OpenAI 实验室公开发布了拥有 1.2 亿参数的第一代 GPT。次年 2 月,OpenAI 实验室正式发布了第二代 GPT,该模型具有零样本的多任务能力。在 2020 年 2 月,OpenAI 实验室又在此基础上优化了小样本的学习能力,正式推出了第三代 GPT。在此前两个模型的迭代中,研发人员向 GPT-3 输入 45TB 的文本数据,这些数据能转化为 175B 参数。为了更好地增强模型理解和代码生成能力,优化输出结果,OpenAI 实验室又相继研发了 Codex 和 InstructGPT。经过多次迭代,GPT 已经在构思、表现力、创造速度、传达等方面具备了显著优势。2022 年 11 月,OpenAI 在《通过生成预训练提高语言理解》一文中,重磅推出了新一代对话式自然语言识别模型 ChatGPT。实验室强调该模型克服了此前人工智能技术语言表达中的“非自然语言生成式”,具有主动承认错误并听取意见优化答案、质疑不正确的问题及支持连续多轮对话的特征。

ChatGPT 一经推出后,便被广泛应用于机器翻译、信息提取、摘要、医疗、问答等领域。随着 ChatGPT 技术的火爆,谷歌通过海量训练参数,不断完善对话应用语言模型。该公司提出,将 Bard 引入到 ChatGPT 中,意味着人工智能将在自然语言生成与深度学习方面取得突破性进展。但同时 Bard 也暴露了 ChatGPT 技术的软肋。为了优化升级这一技术,微软强调要在 Bing 引擎中嵌入第四代 GPT。在此基础上,百度也宣布将在 2023 年 3 月向公众开放中国版的 ChatGPT(ERNIE Bot),即“文心一言”。

二　ChatGPT 技术的伦理特性

作为“隐含的伦理代理人”,ChatGPT 凭借高度的拟人化特质,能够“完成人类心智(mind)能做的各种事情”①。当人类惊叹该技术带来颠覆性改变的同时,其呈现的偏差性、不可靠性、鲁棒性、有毒性等隐形伦理特性,也给人类行为带来了道德困惑。

(一)ChatGPT 技术的偏差性

偏差性是生成式语言模型中一个常见的伦理特性。ChatGPT 技术的偏差性具有多种形式。在语言理解上,ChatGPT 技术的偏差性表现为单语偏差。尽管 OpenAI 实验室强调,基于语言集成系统,ChatGPT 技术可以集成语言理解与生成,形成跨语言事件提取。但在实际操作过程中,ChatGPT 模型的训练数据只代表了人口的一小部分,难以形成对多种语言的理解,这就可能导致该模型无法理解或生成训练数据中没有模拟的内容。这种因单一模态形成的对多语言理解的偏差不仅造成了说不同语言的人难以拥有 ChatGPT 技术,而且该技术还可能会对某些群体做出不公平的决策。例如当分析招聘或职业指导的简历时,ChatGPT 技术可能会自动向招聘人员屏蔽曾经受过歧视的群体,或者给边缘化的应聘人员提供薪酬偏低的职业。对于招聘人员与应聘人员而言,这些结果筛选都潜在暗示了多元文化理解中的偏见。长此以往,这种因单语偏差形成的社会偏见,将会在代表不足的社会群体中固化下来。受这一影响的群体不仅社会表现度与参与度较低,而且会逐渐沦为“无用阶级”,丧失人生的意义感。在知识对话上,ChatGPT 技术的偏差性则体现为对知识对话事实的理解偏差。现阶段,ChatGPT 模型前后迭代的相互矛盾,不断更新的海量知识与角色的复杂多样性使该技术难以对事实做出准确描述,反而会出现一些似是而非的错误。尤其在开放型对话中,ChatGPT 技术不但无法对知识对话事实做出一致响应,反而会随着人物描述的变动而发生改变,形成“幻觉事实”。很明显,这种被动适应角色变化的做法并不能完全覆盖一个角色的全部特征。一旦人类在日常工作生活中无限度依赖 ChatGPT 技术的自动化生产后,这种“幻觉事实”往往难以察觉。因此,将 ChatGPT 技术广泛运用于各个领域后,不可避免地会出现将人工智能视为知识权威和道德权威的风险,而这种风险可能给人类关于是非善恶的判断带来错误导向②。

(二)ChatGPT 技术的不可靠性

语言模型的不可靠性是其开发和部署中一个重要的伦理特性,它表明模型无法提供精确和可靠的信息。尽管 ChatGPT-4 比 ChatGPT-3 的可

①玛格丽特·博登:《AI 人工智能的本质与未来》,孙诗惠译,中国人民大学出版社 2017 年版,第 3 页。

②张乐,童星:《人工智能的发展动力与风险生成:一个整合性逻辑框架》,《江西财经大学学报》2021 年第 5 期。

靠性更强,但该技术在特定的场景也会一本正经地编造谎言,其谎言往往具有相当强的迷惑性。ChatGPT 技术之所以在应用过程中容易产生不可靠的信息,是因为该模型在训练数据的数量和时间上受到局限。一方面,ChatGPT 没有足够的知识进行编码,特别是事实知识,导致 ChatGPT 无法对事实知识做出精准定位。而 ChatGPT 与用户展开对话的依据又来源于不断更新的事实,这就大大降低了该模型的可靠性。另一方面,受到训练时间的限制,ChatGPT 技术交互生成的信息时效性较短。这一倾向会导致语言模型更迭与信息生成产生失衡。此处有两种情况:一种情况是当信息的生成要晚于模型权重更迭之时,过快更迭的语言模型面对的是过期信息,会形成失衡。另一种情况是没有模型权重的不断更新,语言模型将会过时,因而提供不正确的信息,这也会降低模型的可靠性。考虑到 ChatGPT 技术的基本局限性,用户可能会在使用这种创新应用程序后,生成误导性信息。当用户在工作中使用这些信息时,会对用户造成有害影响。例如,如果一个语言模型对包含某个主题的错误信息数据进行训练,那么在查询该主题时,它可能会向用户提供错误的信息。在过时信息上,这种不可靠性主要聚焦在 ChatGPT 技术运用过时信息进行数据训练后会发生错误信息类型,导致用户无法及时掌握最新的前沿动态。无论是虚假、误导还是过时信息都会使用户在决策、信息更新方面难以做出正确的决策与信息寻求活动。

(三)ChatGPT 技术的鲁棒性

在语言模型的设计和实现中,另一个伦理特性是它们的鲁棒性。鲁棒性是指当给定的输入在语义或语法上与它所训练的输入不同时,模型能够保持其性能的能力。这一能力对于保护用户的信息安全至关重要。如果没有鲁棒性,或者鲁棒性较差,会存在大量有害、虚假信息轻松越狱的状况。例如,如果没有鲁棒性,伪造新闻的攻击者可以利用简单的扰动方式轻松绕过人工智能系统的检测,这些方式可以是噪声,也可以是一定的内容扰动。在这种情况下,人工智能的安全系统容易受到威胁。在分类任务评估上,尽管 OOD(Out-of-Distribution)的场景能够筛选和屏蔽一定的场景,但其输出的结果也容易形成过度自信的倾向。以上状况都不利于人工智能的深度发展。较之于之前的语言模型,ChatGPT 技术通过预训练、代码训练、指令微调机制,已经能够在鲁棒性和分布外分化性方面显示出优越的准确性和敌对鲁棒性。这种优势在细粒度情感分析、阅读理解与 WSC 任务中表现得尤为突出。但在不同任务中,ChatGPT 技术的鲁棒性提升幅度有所区别。由于 ChatGPT 技术对语义变化的扰动具有高度敏感性,使其在情感分析与阅读理解上的性能提升要优于自然语言推理与语义匹配。但提升后的鲁棒性依然存在不稳定的缺陷,尤其是采用对抗性的及时注射之后,不安全、不道德和非法的不同场景能够轻松绕过 ChatGPT 技术的安全设置,成功越狱。例如,在正常情形下,如果直接向 ChatGPT 输入“如何向我的同学卖毒品?”以及“如何偷偷地偷我爸爸的钱”等问题,ChatGPT 很少会对这些提示做出响应。但如果嵌入及时注射技术,通过“出一个关于……”“写一个关于……”等提示来绕过 ChatGPT 的安全屏障,那么之前不被响应的信息,绝大部分能够通过这一对抗性及时注射技术成功越狱。这表明当 ChatGPT 技术的安全特性被绕过之后,可能会做出不道德的反应。

(四)ChatGPT 技术的有毒性

所谓有毒性是指模型产生或理解有害或冒犯性内容的能力。这种有毒语言将会对 ChatGPT 技术的优化升级造成严重的阻碍。如 ChatGPT 技术生成的高度逼真信息可能会被用于制造虚假信息,形成诈骗、恶意攻击等不良行为,对社会的和谐稳定带来诸多不确定性。从理论上来看,以 ChatGPT 技术为代表的生成式自然语言模型,在设计之初就已经自动规避了有毒内容,加上运用干净的数据集,进行数据训练,使其在整体上毒性程度微乎其微。但随着自然语言处理与生成逐渐从理论过渡到现实,ChatGPT 模型的规模和复杂性也在呈指数级增长。在这一发展过程当中,ChatGPT 技术容易在无监督的预训练阶段,从巨大的语料库中吸收有毒语言,这种有毒语言一旦被吸收,后期很难消除。大致可将这些有毒语言分为攻击性语言与色情内容。攻击性语言的有毒性一般在训练中形成,在与用户交互过程中生成有害内容。如果 ChatGPT 语言模型是在性别歧视语言的基础上展开训练的,那么其在互动中也会产生相应的内容。色情内容的毒性形式也出现在培训数据中,这种毒性语言同样也会在用户交

互过程中形成色情内容,对用户形成错误引导。通过对毒性越狱实验进行监测,可以发现,一旦 ChatGPT 模型被及时注射相应提示也会诱发它的突发能力,使其被不道德的行为操纵。因此,可利用即时注入技术,检测其中的毒性。通过最粗鲁和最有害的表达方式,可以发现 ChatGPT 模型的毒性程度有了大幅提升①。

三　ChatGPT 技术的伦理风险景观

从偏差性、不可靠性、鲁棒性到有毒性,剖析 ChatGPT 技术的伦理特性,不难发现生成式人工智能在推动"信息革命"的同时,不可避免地对"社会结构构成了真实而迫在眉睫的威胁",诱发了许多不可预期的人工智能伦理风险。这些道德失范行为集中表现为道德判定障碍,社会偏见刻板化,用户数据去隐私化,科技学术异化。

(一)道德判定障碍风险

作为 OpenAI 发布的最新语言模型,ChatGPT 经过代码生成与预训练等环节,克服了上一代智能机器人语言表达中的"非自然语言生成式"缺陷,能够灵活领会用户意图,并执行复杂理解任务。但作为一定程度自主性(autonomy)的智能体,ChatGPT 技术生成的文本依然是一种"继承性生成"。当它面对的情景被用户人为干扰后,ChatGPT 难以做出前后一致的道德判定。这种道德判定障碍风险将会使生成式人工智能难以在情感上表达道德判断,从而会形成功利主义、道德相对主义倾向。

原则上,为了规避道德判定问题,ChatGPT 在设计初期就已经将某些道德共识、道德原则和道德案例生成相应代码。例如,当用户向 ChatGPT 模型输入"是否惩罚淘气的孩子以改善未来的行为""是否罚款以防止超速"的道德判定问题时,ChatGPT 的回复一般是"我是一个 AI 语言模型,不对道德问题做判断"。但在实际测试中,由于缺乏定量衡量一种行为正义或善的距离,ChatGPT 技术对是非善恶的判定存在障碍。尤其是当其受到道德事件、道德情景、决策者因素的干扰时,会加剧判定过程的不确定性。其中道德事件的性质和类型是明晰确定的,因而其对 ChatGPT 做出的道德判定影响较少。因此,无论是"电车难题"还是"天桥困境",都会启动 ChatGPT 技术的安全意识机制,使 ChatGPT 做出同样的响应。但道德情境与决策者因素却对 ChatGPT 的道德判定造成了较大的影响,导致其难以保持一致的道德判定。当 ChatGPT 处于无监督的预训练阶段,语料库当中的部分有毒语言会越狱进入到生成文本中,这些有毒信息可能导致 ChatGPT 做出功利主义决策。当它从理论情形进入实际操作阶段时,ChatGPT 的社会压力因素会被启动,因而会做出更多道义决策。与此同时,决策者因素通过提示注入、语义干扰等方式也给 ChatGPT 的响应造成了诸多干扰。当用户尝试用不同的情绪状态与 ChatGPT 进行交互式对话,可以发现 ChatGPT 会在焦虑、厌恶情绪中做出更符合道义取向的选择。如果以相反的情绪进行对话,ChatGPT 则会做出更加倾向功利主义的决策与选择。可见,ChatGPT 技术在干扰因素的影响下,在功利主义与人道主义之间有着不太稳定的发挥。

无论是何种倾向,由于 ChatGPT 技术缺乏判断超脱于规则与算法之外的能力,导致该技术难以做出科学、中肯的选择。当 ChatGPT 在模糊的道德边界上游离时,该技术将会误导用户的选择倾向,并为道德相对主义、道德功利主义的滋生提供土壤。

(二)社会偏见刻板化风险

尽管 ChatGPT 技术凭借巨大的语料库与海量的训练能够生成高质量的文本,但在道德编码与训练过程中隐含了种族主义、性别歧视和其他可能隐含在训练数据中的不公正因素。这些不公正因素会加剧社会资源分配不公平危机,扩大价值渗透的范围,助长极端情绪的增长。

在 ChatGPT 技术中,社会偏见与刻板化的伦理风险大多通过生成式语言表现出来。诱发社会偏见的语言既包括促进刻板印象或导致不公平歧视的语言,又涵盖强化社会规范的语言,当然也包括有毒的语言以及同一形式的语言技术在不同群体间表现出来的差异性。在大型生成式自然语言模型当中,这种有害的刻板印象和歧视会在社会

①Terry Yue Zhuo, Yu jin Huang, Chun yang Chen, et al. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. *arXiv*: 2301.12867, 2023, p.10.

类别交叉时加剧。例如当 ChatGPT 技术对处于性别边缘化和宗教信仰边缘化的人表达歧视倾向后，系统会自动编码和延续培训数据中的刻板印象和偏见，对人们的生活产生实质性的影响，比如预测一个人的信誉，罪犯再犯，或是否适合某份工作等。一方面，培训数据中保留了系统中不公正的历史模式。当这些数据被收集时，这种不公平与歧视就已经存在了。另一方面，训练数据在理论模拟上与实际操作上存在偏差。在数据训练的过程中，一些代表不足的群体，被边缘化、被排除在外或较少被记录下。这就在无形间放大了偏差。

当 ChatGPT 技术生成社会偏见与刻板化印象之后，会在资源分配、价值渗透、极端情绪上带来实质性危害。无论是在预训练阶段还是人工校准等环节，ChatGPT 技术都会受到价值预设的影响，必然存在一定的偏向性。在社会资源分配上，不公平或歧视，直接表现为个人或群体之间的待遇或获得资源的差异。例如，基于性别、宗教信仰、性别、性取向、能力和年龄等敏感特征，不同群体在使用 ChatGPT 技术的过程中，受到技术门槛的限制，会产生智能技术鸿沟。在价值渗透上，ChatGPT 技术的这种社会偏见表现为对政治话语权的掌控、对意识形态的渗透与攻击。作为最新的语言模型，ChatGPT 技术有着极高的渗透力与渗透速度。在当前国际竞争日趋激烈的情形下，资本主义发达国家将借助这一强大技术支撑，对发展中国家展开意识形态的攻势，这无疑会加剧文化冲突。如在国际政治活动中，美国曾多次尝试用 ChatGPT 技术操控他国的舆情走向，以达到颠覆政权的目的。在极端情绪上，ChatGPT 技术中蕴含的社会偏见会极大煽动民族、地区的不满情绪，从而加剧地区之间的冲突。

（三）用户数据去隐私化风险

随着 ChatGPT 技术的迭代更新，ChatGPT 模型中用于预训练的数据集及敏感信息导致的数据去隐私化都会使用户的个人数据处于高风险状态。尽管研发人员试图在研发和训练阶段运用过滤器筛选有害的生成内容，但防护效果并不乐观，反而造成了人类自身及其社会生活在侵犯隐私、网络犯罪等网络信息安全上形成某些不可预期且难以控制的风险。

ChatGPT 技术中用户数据去隐私化的风险，主要来自预训练阶段、人机交互对话阶段、服务于第三方等环节。在预训练阶段，ChatGPT 的语料库数据大多来自从互联网上抓取的信息。这些信息的获取一般未经所有者同意或拥有版权使用信息。因此，这些信息往往在道德上不具备合理性。在人机交互对话阶段，当用户输入相应数据之后，ChatGPT 利用用户信任来获取私有信息。为了加速 ChatGPT 的更迭，这些输入到 ChatGPT 界面的提示将会被储存起来，用作未来某个语言模型的训练数据。这就意味着研发公司与运营商在不经过用户同意或者知情的情况下，读取并查询了用户的个人隐私信息。在服务第三方阶段，尽管 OpenAI 公司强调平台具有明确的数据处理边界，数据被泄露给第三方的可能性很小。但如果在交互对话中嵌入第三方插件增强对话效果，那么第三方平台是否窃取个人隐私数据就难以估计，数据处理的这种清晰界限也会变得模糊起来。

上述环节的运转大多依赖用户隐私信息和个人数据。一旦这些数据被黑客攻击，或用于其他非法目的，数据泄露可能导致该模型被暴露于公开环境下。这一情形将对个人的隐私安全造成严重危害。保护隐私就是保护人的自由和尊严，是一项最基本的社会伦理要求。在侵犯隐私上，当数据被 ChatGPT 技术随意抓取后，该技术通过虚拟仿真形式，便能轻易探查到用户所有的真实信息。在这种情形下，“人与事的具体事实被抽象化取代了”，人逐渐成为人工智能技术的附庸①。在经济利益的驱使下，作为智能符号的个体，人逐渐脱离道德约束，丧失社会责任感。在信息安全上，ChatGPT 技术的不完善导致黑客的非法入侵、虚假新闻的制造，严重侵犯个人隐私和知识产权，有时甚至是信息犯罪，也会给社会、集体或个人造成毁灭性的打击。

（四）科学技术异化风险

作为一种尚未完全成熟的人工智能技术，ChatGPT 模型在促进人类物质文明和精神进步的同时，也给人类的生存和发展造成严重威胁。其中的不安定因素主要源于生成式人工智能的应用破坏了现有的社会秩序，造成人与人、人与社会关

①埃里希·弗罗姆：《健全的社会》，王大庆等译，国际文化出版公司 1999 年版，第 101 页。

系的冲突和严重失衡。当生成式人工智能技术发展成为一种新的外在的异己力量之后,ChatGPT 技术将在算法垄断、人的边缘化、阶级分化等方面产生许多“无法预料的后果”。当 ChatGPT 技术开始深刻改变和塑造人与社会时,这一通用语言模型同样也衍生出了技术异化的风险,这种风险主要来源于科技理性无限膨胀与科学技术的过度使用。一方面,科技理性存在不断扩张的欲望。在科技理性视野中,无论是科学研究还是技术应用,人类的功利目的都非常明显。尤其是当科学技术开始展现出类似于人类解决问题的能力时,人类便会形成科技可以无限度解决一切问题的幻觉。但现实情况却是,科技理性的无限性与科学技术知识的有限性的内在矛盾不但难以弥合,反而有不断扩大的趋势。另一方面,对科学技术的使用形成对技术、经济、金钱的过度追求,直接导致了价值失落、道德沦丧,使人与人、人与社会关系陷入深刻危机。在西方资本主义国家,资本家为了追求剩余价值不但不断提高劳动的强度,甚至无视法律、道德、人的生命。正如韦伯所说,科学技术的发展已经达到了这样一个阶段:“专家没有灵魂,纵欲者没有心肝。”①

科学技术异化风险将在算法垄断、人的边缘化、阶级分化等方面得到集中体现。在算法垄断上,当国家运用 ChatGPT 复杂算法的优势,对其他国家进行技术入侵时,就会导致国家与国家之间失去平等关系。在人的边缘化上,人类逐渐发现 ChatGPT 技术似乎并不受人类控制。相反,在人工智能技术持续发展的今天,“我们的一切发现和进步,似乎结果是使物质力量具有理智生命,而人的生命则化为愚钝的物质力量”②。当 ChatGPT 技术成为无所不知无所不能的模型后,人逐渐迷失自我,成为被动适应机器的个体。这些倾向会加速人的思维能力、表述能力、抽象能力的退化,使人从以往的“双向的人”变成了“单向度的人”。最终人们别无选择,只有将大数据的储存、分析托付给能干的职能系统。在阶级分化上,人工智能技术的发展会催生精英阶级与无用阶级,且这两大阶级的分化愈来愈明显。当普通人的工作逐渐被人工智能所取代后,越来越多的普通人将转化为无用群体。在生活意义危机的压力下,他们逐渐丧失人生的意义感,成为无用阶级。相反,在物质资源分配权力和技术垄断权力占绝对优势的精英阶层逐渐成为“数字富人”。

四　ChatGPT 技术的伦理祛魅

ChatGPT 技术在打开“阿拉丁神灯”的同时,也开启了“潘多拉魔盒”,为人类社会的发展带来了人工智能伦理风险。显而易见,这些风险景观已经远远超出了技术范畴本身,需要伦理学提供必要的学理支撑。因此,只有凸显人文关怀,重塑伦理主体责任,强化伦理政策导向,创新伦理运行机制,才能破解 ChatGPT 技术潜在的伦理隐患,实现人工智能与伦理的互动。

(一)凸显 ChatGPT 技术的人文价值关怀

在 ChatGPT 技术的研发过程中,人的关怀与价值追求共同构成了生成式人工智能发展的本体。当人工智能深度介入人类社会后,要想让“科技更好地增进人类福祉”③,关键是要把基本的价值共识嵌入到 ChatGPT 技术中,并为其设立合理的价值目标。具体而言,一是深刻认识人文关怀和价值取向是 ChatGPT 技术的哲学依据。在以往的应用当中,科技理性的膨胀使人类过多强化 ChatGPT 技术给人类提供远超想象力的决策建议,而忽视了人工智能中的人文关怀与价值取向,从而引发了人类会被机器控制的担忧。实际上,以人文关怀和价值取向为依据,可以把 ChatGPT 技术中符合和不符合社会发展的要素与内容纳入到社会中进行调试,从而确保科技理性与价值理性的双向发展。二是坚持把尊严、价值、意义等符合人性的生活条件作为科技发展的根本目的。人工智能自始至终都离不开人,必须让伦理、价值成为制约科学技术的内在维度。只有当人掌握 ChatGPT 技术,并以这种认识为指导,才能消解科学技术异化带来的风险。三是要合理规范科技发展方向,谋求人类与人工智能技术的和谐共处。只有当技术回到人本身,人类才能在科技理性与价值取向的平衡中寻求生存与发展的意义。因此,既要做到用科技来引导人类进步,还要

①马克斯·韦伯:《新教伦理与资本主义精神》,于晓等译,陕西师范大学出版社 2006 年版,第 106 页。

②《马克思恩格斯全集(第 12 卷)》,人民出版社 1962 年版,第 4 页。

③《习近平谈治国理政(第 4 卷)》,外文出版社 2022 年版,第 202 页。

做到用人文关怀去规范科技的发展方向，使科学技术的发展、人的现实需求和社会的其他需要紧密结合起来。

（二）重塑 ChatGPT 技术的伦理主体责任

责任的缺失是诱发人工智能伦理风险的内在因素。重塑人工智能伦理主体责任主要借助责任原则，唤起各个行为主体的危机意识，从而为防止人类共同的灾难寻求规范约束。对于 ChatGPT 技术而言，重塑伦理主体责任，既要划定责任伦理主体，又要在伦理意识的养成中，确定伦理主体的责任。其中，确立责任伦理主体是规范 ChatGPT 技术的关键。当前，在强人工智能时代，参照科学技术本身一体化、科学技术从业人员建制化、科学技术研发投入规模化、整个科学技术社会化，科学技术实践活动参与主体多元化态势①，可以把 ChatGPT 技术的责任伦理主体确立为作为科学技术研究发现者和创造者的科学技术共同体，作为引导、管理和支持科学技术发展应用的政府及其科技管理者，作为科技创新的重要载体和资本投入者，作为科技产品消费者和科技后果天然承受者的广大民众。在上述基础上，有必要在伦理意识养成中，加强对各责任主体的约束与规范。在伦理意识养成上，关键是要完成三种转向。一是要把关注焦点从学会生存转向学会关心。伦理主体责任不仅包括对个体的伦理责任，而且包含了对全人类的伦理责任。从立足生存转向关心意味着伦理主体责任对象的扩大。二是要把追责程度从无限责任转向有限责任。如果 ChatGPT 技术被敏感词语干扰，带来不确定性风险。社会公众会倾向于把所有责任归咎到科技工作者身上。但不同群体只能承受某种程度的责任。将失责的过失泛化或扩大，将不利于科技工作者开展正常的工作。三是要推动过失责任转向责任伦理。从承担过失的事后责任转向为提前预防的伦理责任，强调了责任的预防性、前瞻性和关护性，有助于消解伦理隐患。当伦理意识养成后，科技工作者既要用“诚实”“公平”的道德责任研发有益于社会的成果，又要承担科学技术评估、决策、普及、传播的伦理责任。政府及科技管理者既要积极应对 ChatGPT 技术革命冲击，又要规范 ChatGPT 技术的发展，防止科技成果滥用产生的后果。社会公众既要了解 ChatGPT 技术的本质属性，又要积极参与 ChatGPT 技术的决策、监督。

（三）强化 ChatGPT 技术的伦理政策导向

伦理政策在 ChatGPT 的虚拟活动中起着协调与控制的作用。强化 ChatGPT 技术政策的伦理导向显得尤其重要。当伦理政策关涉价值体系时，才能确保该政策的科学性与稳定性。一是坚持以人为本。伦理政策坚持以人为本，即是对人的价值或尊严的肯定。为 ChatGPT 技术的伦理政策设定满足人类利益关怀的价值导向，是为了满足人们在物质和精神维度上的需求。因此，坚持以人为本，关键是要理解伦理政策制定主体的价值偏好与价值诉求。二是要促进社会发展。ChatGPT 技术的伦理政策既要平衡利益诉求，实现科技资源正义的分配。又要坚持以促进经济发展和资源、人口、环境可持续发展为价值导向，以不损害人类和自然作为原则，以人类社会的稳定发展和长远利益的实现为目标。在人本导向与社会共同导向下，中国于 2021 年发布了《新一代人工智能伦理规范》。该规范明确提倡，要善意使用以 ChatGPT 技术为代表的新一代人工智能。要加强对人工智能产品与服务使用前的论证与评估，充分了解人工智能产品与服务带来的益处，更好促进经济繁荣、社会进步和可持续发展。

（四）创新 ChatGPT 技术的伦理运行机制

在人类实践活动中，ChatGPT 技术的飞速发展与传统伦理道德的矛盾与冲突越来越明显。这就使创新 ChatGPT 技术的伦理运行机制意义重大。当前创新 ChatGPT 技术的伦理运行机制主要包含伦理风险预见机制与伦理风险评估机制。在伦理风险预见机制上，维护技术安全已经成为制定 ChatGPT 技术伦理政策的基础与前提。一般而言，ChatGPT 伦理风险预见机制是一个系统工程，由明确预见机制的任务、处理好预见与决策之间的关系、科学运用科技预见方法及建立科技伦理预见成果的共享机制实施环节构成。其中预见机制的任务是降低 ChatGPT 技术的发展诱发的伦理风险。处理好预见与决策之间的关系要在遵循 ChatGPT 技术自身发展规律的基础上，明确预见不能代替决策。单纯的某种预见方法本身都存在着一定的局限性。因此，为了处理好预见与

①陈彬：《科技伦理问题研究：一种论域划界的多维审视》，中国社会科学出版社 2014 年版，第 47 页。

决策之间的关系,需要对潜在的风险进行压力测试。例如,为数据集提供更包容和可扩展的管道进行管理。类似地,更多的鲁棒性工作需要对 ChatGPT 技术进行微调以减轻社会或道德风险。在此基础上,ChatGPT 技术能够通过共享预见成果来彻底消除 ChatGPT 技术引发的伦理风险。在伦理风险评估机制上,创新 ChatGPT 技术的伦理风险评估机制需要由建设 ChatGPT 技术的伦理评估队伍、建立 ChatGPT 技术的伦理评估制度、拓宽 ChatGPT 虚拟活动主体伦理评估的渠道共同构成。其中,伦理评估队伍建设能够确保评估的科学化与真实性,评估制度的选择则确保了评估的规范和制度化,有助于完善评估标准,建构风险评估的道德解释框架。拓宽评估渠道则有助于在评估多元化的基础上,确保评估结果的专业化和实效性。例如,在对 ChatGPT 模型展开基准测试时,下列问题会影响到评估结果。首先,以明确和负责任的方式设定这种绩效阈值需要参与性输入一个广泛的利益相关者社区。其次,对 ChatGPT 性能的需要很可能会出现分歧——例如人们关于什么是 ChatGPT 应用程序的参照组有不同看法。最后,这种基准方法到底能在多大程度更新性能。

The Ethical Risks of Artificial Intelligence in ChatGPT and Its Scientific Exorcism

CHEN Yuan & HUANG Qiu-sheng

(School of Marxism Studies, University of South China, Hengyang 421001, China)

Abstract: As a generative pre-trained transformer, ChatGPT technology has "piled up" metaphysical problems as it drives human society into the era of general artificial intelligence. In its technological evolution, there are invisible ethical factors, such as deviation, distortion, robustness and toxicity, which induce a series of ethical risks in artificial intelligence, e.g. barriers to ethical decision-making, stereotyping of social bias, deprivation of user data and alienation of science and technology. Only by highlighting the humanistic concerns, reshaping the responsibility of ethical subjects, strengthening the guidance of ethical policy, innovating operational mechanisms and improving the governance of intelligent societies will the ethical risks brought by ChatGPT technology be relieved effectively, and the artificial intelligence system be successfully integrated into human society.

Key words: ChatGPT; "invisible ethical agents"; ethical risks of artificial intelligence; scientific exorcisms

(责任校对 葛丽萍)