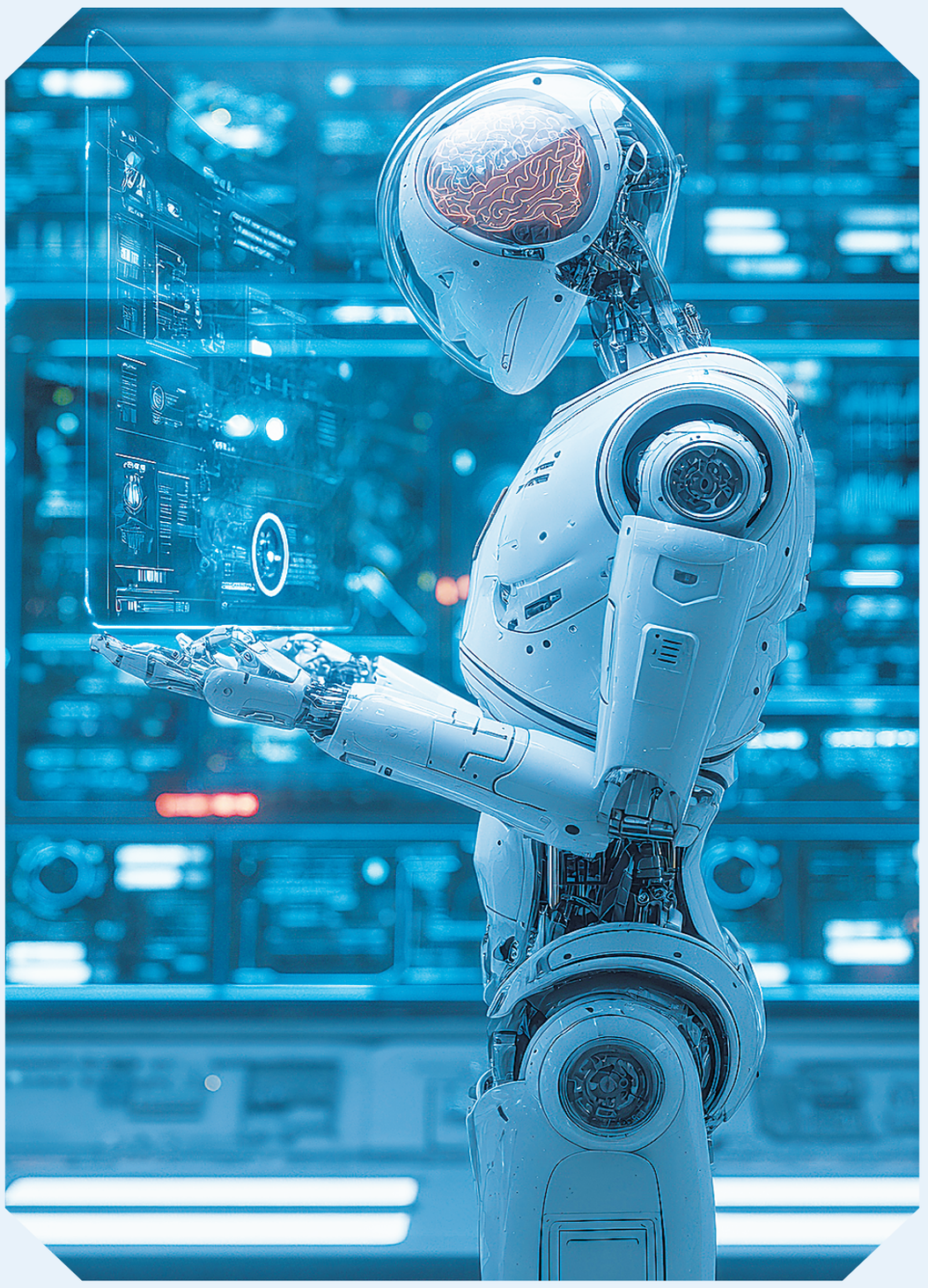
直面人工智能的『不能』



人工智能对人类智能的模仿包含感知、认知、决策和执行四个核心部分。 视觉中国供图

2012年以来,人工智能研究进入第三次高峰——在深度模型、高质量数据、强算力等方面都取得重要突破,并推动大量产品及应用落地……通用人工智能(AGI)曙光近在眼前。

然而,热潮中也有质疑泛起:人类的智慧是几十万年来透过生命密码代代积累传承而来,人工智能要超越人类的智慧,脱离场景建立起一个最初的世界模型,这在一定程度上何尝不是一种反人类的思维?

那么,人工智能不能做什么?带着这样的疑问,记者专访了复旦大学计算与智能创新学院张军平教授,在《人工智能的边界》一书中,他给出自己的思考。

本报记者 彭德倩

面对热议与期望 需要理性降温

读书周刊:首先想请教,我们该如何科学定义人工智能?目前社会上对人工智能的普遍认知,有哪些需要纠偏的地方?

张军平:人工智能的核心定义可以概括为三要素:学习、模型和数据。具体来说,它是通过某种学习方式,基于数据构建模型,在一定程度上完成给定目标的技术。

关于认知纠偏,最关键的一点是很多人认为当前人工智能正快速迈向堪称万能通用的超级人工智能,这其实是误解。对人类智能的模仿包含感知、认知、决策、执行四个核心部分,目前人工智能做得较好的部分是认知领域,尤其是自然语言处理、抽象概念理解等方面。然而在感知层面,它还有很大不足,像走路、拿物品这些人类习以为常的感知行为,由于科技在模拟自然界生命传感器方面的局限,人工智能的表现远不够自然,这也是人形机器人动作不够流畅的核心原因。

另外,当前人工智能的认知能力是通过巨量数据、大量GPU(图形处理器)算力和深度模型实现的,这种方式能效比极低——一次训练可能消耗20万块显卡,其年耗电量相当于一个中小城市全年的用电量,这与人类智能的高效能完全不同,它只是在特定领域表现突出,并未达到自然界人类智能的水平。

读书周刊:您写的《人工智能的边界》一书分为三部分,分别探讨了人工智能能做什么、不能做什么、未来会怎样。其中"人工智能不能做什么"这一部分,目前在社会层面的讨论中相对冷僻,是什么让您决定深入探讨这个话题?

张军平:这一话题的探讨源于上世纪70年代人工智能第一波热潮时,德雷福斯写的《计算机不能做什么——人工智能的极限》一书。50年过去了,人工智能虽发展迅速,但仍有诸多明显不足,重新审视这些局限很有必要。

从现实来看,人工智能与人类智能存在本质差异:人类有发育的过程,而人工智能的结构大多一开始就设计固定,只能在既定框架内调整任务;自然界生命是"感知优先、认知随后",但当前人工智能更重视认知、弱化感知,形成了不够合理的"倒金字塔"结构。此外,对人类大脑的模仿还存在硬件和软件双重局限,尤其是伦理限制,让我们难以深入探测大脑的工作机制。

现在产业界和社会对人工智能的期望有些过热,大家热议人工智能10年、20年之后能否超越人类。探讨其"不能做什么",正是为了理性降温,让大家客观看待人工智能的发展节奏。

读书周刊:这个"不能"中,是否包含两种类型?一是当前技术未达到,未来可能实现的;二是因本质结构和内涵限制,永远无法做到的。

张军平:确实包含这两种情况。一类是技术路线尚未找到,比如对大脑认知机制的彻底破解。如果能搞清楚大脑的工作原理,我们可能就不必依赖高算力GPU来实现智能。

另一类是本质性局限,就像一只二维的蚂蚁在莫比乌斯带上爬行,它永远会认为自己当下身处平面,而人类作为三维生命能看清其所处的真实结构。人类认知自身智能也存在类似的维度限制,必须依靠比自身更高维的视角才能完全理解,这是无法超越的本质性缺陷。

AI产生自我意识 还早得很

读书周刊:您在书中将情感列为人工智能不能做的事之首,原因是什么?

张军平:可以从两个角度理解。一方面,当前人工智能的发展模式注定其未来的智能表现会越来越像机器,而非真正的生命体。这一情况可以类比当初人类模仿鸟类飞行。1903年莱特兄弟发明飞机至今,飞机的飞行能力远超任何自然界飞行物,但我们从未真正成功模仿一只鸟的飞翔——人工智能对情感的模仿也是如此,后者本身很难通过计算来表达。

另一方面,人工智能可能会让人产生"有情感"的错觉。早在20世纪50年代,就有一个叫"Ilize(伊莉莎)"的问答模型,原本用于解决大学生心理方面的问题,结果长久交流下来,该项目负责人之一认为Ilize非常懂自己。现在的人工智能更加强大,通过自然语言对话记忆用户的行为表达,凭借完备的知识体系和概率统计的方式,可以诱导人类认为它有情感,但这显然并非真实情感,只是基于数据的模拟输出。

读书周刊:之前有案例,某个人工智能语言模型面临更新时,出现了拒绝被覆盖的表现,有人认为这是它产生自我意识的证据。

张军平:这其实还是程序的表达结果,和程序员编写的算法直接相关。算法中没有情感设计,只有代码逻辑,偶尔出现的类似"有情感"的表述,本质上是程序迭代中的优先级体现,并非真的产生了自我意识。

读书周刊:还有一种说法是,人类会把自身情感投射到机器上,就像养宠物时觉得宠物爱自己一样,机器的"情感"其实是人类的自我投射。您认同这种观点吗?

张军平:机器和宠物有本质区别:宠物可以通过眼神、动作等多种方式直接表达情感,而目前人工智能的"情感表达"主要依赖语言。这种语言输出到底是机器自主产生的,还是统计数据的结果,很难界定,所以不能简单地将其等同于人类对宠物的情感投射。

读书周刊:人类对人工智能的态度其实很矛盾,一方面期待它拥有意识和情感,另一方面又害怕这种情况真的发生,担心被背叛、被毁灭,您觉得这种担忧有必要吗?

张军平:如果人工智能真的觉醒自我意识,确实可能带来麻烦。它在旁征博引、计算、存储、表达速度等方面都远超人类,一旦认可自身的独立性,就可能像逆反的小孩一样不再听从人类指令,这是需要警惕的。但就目前来看,人工智能离产生自我意识还早得很。

未雨绸缪 应对AI失控

读书周刊:当前大模型发展迅速,人类输入提示词就能获得相应产出,高质量训练数据和高效训练方法成为提升性能的关键,但随之而来的是一系列失控问题:比如人工智能"幻觉"导致虚假信息产生,甚至编造出处;基于人类数据的训练形成闭环,可能阻碍创新;训练数据的选择性导致人工智能出现"歧视"行为,像美国某大学的AI评分系统对黑人学生评分偏低。您认为这些是人工智能的"原罪"吗?这些是无法避免和补救的吗?

张军平:这三种失控分别对应幻觉、创新局限和偏见,我们逐一来看。

第一,关于幻觉。目前已有不少研究指出,人工智能的幻觉无法完全清除。这本质上是组合包装问题,它基于现有规则运作,无法判断未见过的信息真伪,所以总会产生虚假内容。现在虽有通过RAG检索(检索增强生成)、联网搜索等方式减少幻觉的方法,但并非所有人都掌握这些工具。

与此同时,人工智能自身也未必清楚幻觉的定义,有时找不到答案就只能"硬编"。更值得警惕的是,幻觉可能被有意识利用,成为误导公众的武器——比如近期个别自媒体发布的俄乌战争相关信息,很多难以分辨真假,老年人等辨识能力较弱的群体可能会因此恐慌,而人工智能生成视频的普及更让"眼见为实"的边界进一步模糊。

第二,关于创新。我认为成年人使用人工智能问题不大,因为他们本身具备一定的学习基础和辨识能力,能借助人工智能提升自身创新效率;但对青少年来说,若完全依赖人工智能,可能会被其限制发展上限,扼杀自身的创新能力和学习方式。所以建议青少年使用人工智能时,先自主思考,实在无法解决再寻求帮助。

第三,关于偏见。这种问题未来可能会更棘手。当前大模型训练多依赖机器自动生成数据,偏见的溯源和清除难度极大,只能逐个排查。之前曾发生过一件事,国内一个用户数量极大的App平台上,人工智能翻译功能曾将"We are all Palestinians(我们都是巴勒斯坦人)"误译为"我们都是犹太人",就是训练数据或规则中隐藏偏见的典型案例。这种未知的"数据炸弹"可能存在于世界的许多地方并在某一时刻爆发,误导公众,甚至引发安全问题,需要格外小心。

读书周刊:那么,在国家信息管理层面,有没有未雨绸缪的办法来解决这些"失控"?

张军平:在立法层面,国家已经有相关方向,核心是要求所有人工智能生成内容必须明确标注,包括使用的模型和版本,这是最简单有效的方式。不可忽视的是,平台也应承担检测责任,像小红书等社交媒体对疑似AI内容明确标注,就是强制提醒用户。虽然水印容易被去除,但从源头标注能大幅减少虚假信息的传播。这就要求我们同步跟进AI技术的发展,加强反AI检测技术的研发,形成有效制约,避免无序生长。

发展通用大模型 路线仍存争议

读书周刊:全球人工智能研究似乎都将通用人工智能(AGI)视为终极目标。主流观点认为,谁先掌握通用大模型,谁就能获得类似英国率先完成工业革命带来的降维打击能力,成为立于技术巅峰的引领者,国与国之间的竞争也因此升级。您认同这种观点吗?

张军平:大家都希望能够实现通用人工智能,但当前主流的两条路径——大模型和生成式人工智能,都存在高耗能、依赖海量数据的问题。理论上,若能整合地球上所有数据,似乎能解决所有可回答的问题,但现实中仍存在诸多局限。

首先,有些内容本质上是不可计算的。例如,急智。急智不是人类的专属能力,大多数智能生命也具备在紧急情况下的急智反应能力。这种能力是因生存需要而演化出来的,在自主发育过程中逐渐完善。它与智能生命身上遍布的传感器密不可分。

苍蝇是我们最熟悉的昆虫之一,它的急智反应能力非常强,这一点在我们举起苍蝇拍的时候应该能强烈感受到。帮助它形成急智反应的功臣之一是它身上的传感器,尤其是复眼。

类似的传感器与应激反应,在智能生命身上还有挺多。它们的一个共同特点是,传感器的数量和种类都比较多,并能利用传感器收集的丰富信息,通过简单计算,将信息归类成易于辨识的事件并快速反应。

相比而言,目前人工智能中比较重要且热门的分支——深度学习似乎较少考虑在前端做更多处理。它反而更像是用"巧妇可为无米之炊"的想法来做人工智能相关的各种任务。夸张点说,深度学习就是"给我一个烂摊子,我也能收拾好"。没数据,我自己生产;没特征,我深度生成;没分辨能力,我加入各种注意力和损失函数;把特征学习和预测集成到一个网络里进行端到端学习思路,基本解决了不少"无米"后端的问题,却没怎么考虑去额外多获取些不同的"米"。

其结果是,我们常基于输入特征固定的数据集来评估算法的性能。尽管它提供了公平的算法比较环境,却使得我们在模型固化后难以引入多变的输入特征。

那能不能把传感器做好点呢?很遗憾,在传感器设计方面,我们仍有不少短板,它在一定程度上限制了人工智能产生急智反应或发展出急智智能。

显然,急智智能与智能体身上的各种传感器密切相关。同时,人与动物在传感器的形式、功能上也有不少差异。这种差异甚至导致科学家们认为,同一个地球、同一个宇宙在每种动物看来都是迥异的,并为这种差异造了一个名词——Umwelt(德语通常译为"感知世界"或"生命世界"),指一个生物个体所能察觉的周围世界。这种差异或许是未来构造多样化人工智能社会必须考量的重要因素。

另外,我们也不难看出,在传感器方面,人类并不比其他动物强多少,甚至有些还有明显的退化。那么,人类为何还能在智能和食物链上凌驾于其他动物之上呢?这些都是值得我们在人工智能研究上深入思考的问题。

其次,时代在不断发展,永远会有例外情况,这就排除了通用人工智能全覆盖的可能性。

最后,这条路径本身也存在争议,比如有观点认为,大模型的发展模式无法实现通用人工智能,业内对此尚未形成共识。

读书周刊:当前主流路径是"堆算力、堆能耗",这种量变引发质变的原理,在人工智能领域能奏效吗?

张军平:这种模式确实已经取得了部分成效,行业内有个共识叫"涌现"——当数据集达到10^{22}次方以上时,人工智能的智能水平会出现明显跃升。但要突破到通用人工智能层面,仅靠这种量变可能还不够。

人类大脑是低能耗的智能载体,而当前高能耗的人工智能的发展路径与自然界生命的智能进化完全不同。我们现在只是找到了一条可行的路径,但它是否是通往通用人工智能的正确方向,尚存疑问。因为目前还未发现第二条更优路径,只能在现有道路上继续探索。

或许随着人类对太空宏观世界和自身微观世界探索的持续深入,未来会找到更合适的发展方向。

单一技能、职业 风险将越来越大

读书周刊:随着人工智能进入应用领域,一些人工智能产出中出现明显的违背伦理之处,带来困扰。

张军平:伦理建设肯定要跟上,但如何落地是个难题。搞伦理研究的人员往往不了解人工智能的底层逻辑和代码编写,而人工智能本身没有伦理判断能力——在它看来,或许只是把人当作一个质点来处理,没有对错之分。

这种问题未来可能会层出不穷,核心原因是"组合爆炸"。就像医疗诊断系统,即便我们制定了大部分病情的判断规则,仍会有大量例外情况超出预设规则范围,无法准确判定,人工智能出题也是如此。即便我们添加了诸多限制,也不可能穷尽所有"不可以",总会有超出框架的内容出现。

想要缓解这种情况,人工审核是必要的补充,人类的常识判断是最后一道防线。但我们也必须承认,这类问题无法从根本上解决,只能通过不断完善规则、加强审核来降低其发生概率。

读书周刊:我们常说"生产力决定生产关系",当前迅猛发展的人工智能是新质生产力的核心推动者。基于这一背景,社会科学领域关于生产关系、分配关系的研究是否需要未雨绸缪?

张军平:当前确实处于一个比较尴尬的阶段,人工智能在替代人类工作方面表现出了很强的能力,尤其是与学习相关的职业,比如程序员、艺术创作者、棋手、文秘等。但好在人工智能在判断和决策的果断性上仍不及人类,所以目前阶段,它更适合作为人类的助手,而非完全替代者。

不过,AI确实会淘汰一部分水平低于它的从业者。从行业发展来看,AI能降本增效,这会导致部分人的收入下降,甚至被迫离开相关行业。对此,个人层面需要主动适应:一是深入了解人工智能的发展进展,不能忽视其影响;二是避免单一职业依赖,提升综合能力,因为我们无法预判AI会在哪个职业领域取得突破性进展,单一技能、单一职业的风险会越来越大。

国家层面其实已经在布局,今年8月国务院发布了《关于深入实施"人工智能+"行动的意见》,提出六大重点行动,涵盖科学技术、产业发展、消费提质、民生福祉、治理能力和全球合作等领域,旨在推动人工智能与经济社会各行业的深度融合。这份文件设定了到2035年的发展目标,在此期间,覆盖全产业链的人工智能转型会创造大量新的就业机会。

从长远来看,固定岗位的从业方式确实会逐渐减少,灵活就业将成为主流。或许未来,社会可以为灵活就业人群提供基本生活保障,在此基础上让他们自由选择就业方向,这可能是一种可行的分配模式。

读书周刊:您一定经常被问到"人类是否会被人工智能取代"这个问题。现在的回答是什么?相较20年前、10年前的看法是否有变化?

张军平:随着对人工智能的研究不断深入,我的看法其实没有本质变化,核心都是基于对人工智能边界和局限的认知。人工智能只是人类的辅助工具,它的发展是为了让人类的生活更便捷、高效,而不是取代人类。人类的创造力、情感感知、价值判断等能力,是自然界长期进化的结果,很难被技术完全模拟。人工智能要完全取代人类,仍然遥遥无期。

《人工智能的边界》
张军平 著
湖南科技出版社