# 人工智能“革命”的“近忧”和“远虑”
## ——一种伦理学和存在论的分析

□ 赵汀阳

中国社会科学院 哲学研究所 北京 100732

### 一、人工智能的“近忧”

尽管有些预言家(例如库兹韦尔)相信达到“存在升级”的人工智能“奇点”已经胜利在望,但更多的科学家认为“奇点”仍然是比较遥远的事情,潜在可能尚未在望,因为许多根本的技术难点仍然不得要领,特别是尚未真正了解思维的本质、机制和运作方式,所以无从断言其到来。在此,我把能够形成“存在升级”的人工智能看作属于“远虑”的知识论和存在论问题,而把将在近年内确定能够实现的人工智能看作属于“近忧”的伦理学问题,这一讨论也将由近及远来展开。作为“近忧”,人工智能的技术应用非常可能面临以下伦理学问题。

其一,自动智能驾驶悖论。这是近年来引起普遍关注的一个实际难题。假如人工智能的自动汽车(目前的技术只是无人驾驶汽车,尚未达到完全自主智能的汽车)在路上遇到突然违规的行人,是保护乘车人还是行人?这似乎很难做到两全其美,于是形成了一个两难选择。严格地说,这是人的悖论,不是机器的悖论。机器只是遵循规则而已,问题在于我们不知道应该为自动汽车选定什么样的规则。这个悖论只是人工智能可能带来的技术应用难题的一个象征性的代表,类似的悖论也许会有很多。此类悖论具有一个通用难点,即当人工智能成为人类的行为代理人,我们就需要为之设置一个“周全的”行为程序,而这正是人类自己的局限性。事实上,人类能够做出许多伟大的事情,却从来没有做过真正周全的事情。这也正是之所以存在那么多哲学问题的一个原因。我们习惯于百思不得其解。

其二,失业问题。这是赫拉利在《未来简史》里提出的问题,即人工智能的大量应用必定导致大量失业。这个迫在眉睫的问题也已经得到广泛重视和讨论,但目前想象的普遍社会福利政策(比如国民基本收入方案)其实并没有正面回答失业问题,而只是另外回答了收入和分配问题。失业问题的要害之处不在于如何合理分配收入(这是能够解决的问题),而在于生活意义的消失。无事可做的人能够做什么?以什么事情去度过时间?是把一生浪费在电脑游戏、影视作品和闲聊上吗?我们有必要来反思劳动的意义。除了作为生存手段的“硬”意义,劳动(包括体力劳动和智力劳动)还有不可或缺的“软”意义:劳动提供了“生活内容”,以哲学概念来说,它是有意义的“经验”,即接触事物和人物的经验。与事物和人物打交道的经验充满复杂的语境、情节、细节、故事和感受,经验的复杂性和特殊性正是生活意义的构成成分,也是生活值得言说、交流和分享而且永远说不完的缘由,是生活之所以构成值得反复思考的问题的理由。假如失去了劳动,生活就失去了大部分内容,甚至无可言说。这里我们也许可以想象一种“人工智能的共产主义”,它大概满足这样的条件:人工智能创造大量财富并且免除了大量人力劳动,同时存在着落实到每个人的普遍高福利的社会分配。那么,按照共产主义的乐园逻辑,在摆脱了被迫的劳动之后,劳动作为人的本质就得以显现,劳动不再是痛苦的而成为人们的第一需要,人们自愿劳动,并且在劳动之余从事反思性的“批判”。然而问题在于,在人工智能条件下,即使自愿追求劳动也已经没有太多事情可做,那么,非常可能的情况是,当人们失去劳动,又有了普遍福利时,“批判”也随之失去意义。显然,假如一切需求问题都解决了,人们皆大欢喜,也就没有留下需要批判或值得批判的问题了。这里可以看到一种维特根斯坦式的现象:许多问题的解决并非有了答案,而是问题本身消失了。在欲望满足之后失去意义,或者说,在幸福中失去幸福,这非常可能是一个后劳动时代的悖论。也许我们可以抱怨人心不足、人性矫情,但此类抱怨于事无补。无论如何,人工智能导致的大量失业只是表面问题,真正严重的实质问题是失去劳动会使人失去价值,使生活失去意义,从而导致人的非人化。在技术进步高奏幸福凯歌的现代时期,人

们乐于想象技术进步是对人的解放,但情况并非如此,技术进步并不是人获得解放而回归自然的机会,结果反而可能是人的异化。马克思似乎没有预料到高科技高福利的全面解放很可能适得其反地导致人的本质异化,即失去劳动机会或者人工劳动失去意义会导致人的存在迷惑。假如未来人的生活就是在苦苦思考何以度日,那将是最具反讽性的生活悖论。

其三,人对人关系的异化。假如人工智能发达到不仅提供大多数劳动,而且提供一切生活服务,就非常可能导致人的深度异化,即人与人关系的异化。与个体人失去劳动的异化相比,人对人关系的异化更为危险。当人工智能成为万能技术系统而为人类提供全方位的服务,一切需求皆由技术来满足,那么,一切事情的意义也将由技术系统来定义,每个人就只需要技术系统而不再需要他人,人对于人将成为冗余物,人再也无须与他人打交道,其结果必然是,人不再是人的生活意义的分享者,人对于人失去了意义,于是人对人也就失去了兴趣。这就是人的深度异化,不仅是存在的迷茫,而且是非人化的存在。我们知道,自从人成为人以来,人的意义和生活的意义都是在人与人的关系中被定义的。假如人对于人失去了意义,生活的意义又能够发生在哪里、落实在哪里呢?假如人不再需要他人,换句话说,假如每个人都不再被他人所需要,那么生活的意义又在哪里?也许对未来的疑问总是受限于我们对生活的传统理解,因而有保守主义之嫌。那么,如果以充分开放的激进态度来面对这个问题,又能给出什么样的价值解释呢?这恐怕仍然是个难以回答的疑问。一切以技术为准的生活肯定是我们目前无法理解的生活,我们尚未发现它可能产生的意义,只能看见我们所能够理解的生活意义在流失。人类生活的意义和人的概念是在数千年的传统(包括经验、情感、文学、宗教、思想的传统)中建构并积累起来的,假如抛弃人的文化传统,技术系统能够建构起足够丰富的另一种文化吗?能够定义另一种足以解释幸福的价值观吗?我们无法预料,只能深怀疑虑。

其四,人工智能武器。要说人工智能的何种"近忧"最为危险,恐怕莫过于人工智能武器,它甚至比核武器还要危险得多,其危险性就在于人工智能武器将使战争变成无须赌命的游戏。显然,只有必须赌命的威胁才能减少战争,一旦智能武器可以代替人进行战争,人不再需要亲身涉险,人们恐怕也就无所畏惧了,懦夫都会变成勇士而特别敢于发动战争。更进一步说,假如人工智能将来获得自我意识——这已属于"远虑"了——人工智能武器就很可能成为人类自作自受的掘墓人。因此,人类无论如何必须禁止人工智能使用武器的能力,至少高能武器(核武器、激光武器、生化武器等)不能交给人工智能,而必须永远属于与人工智能隔绝的、由人操作的另一个系统,即一个与人工智能无法通用的技术系统。由人类全权控制高能武器,不仅是为了减少战争,而且也是为了必要时能够摧毁人工智能系统。也就是说,即使人类一定要发展人工智能,也必须把武器的使用权和使用能力留给人类自己,必须保证人工智能无法操作武器系统,否则人类的末日就可能不仅仅出现在科幻片中了。

## 二、人工智能的"远虑"

尽管具有自我意识的超级人工智能的出现可能尚有时日,但我们也有理由未雨绸缪。我们之所以有必要杞人忧天,是因为人工智能可能导致的"变天"将是无可补救的人类终结,至少也是人类历史的终结。但愿超级人工智能最后被证明只是危言耸听。

可以肯定,人工智能有希望给予人类用之不竭的技术帮助和巨大的经济福利,但太好的事情就可能会有始料未及的副作用,甚至可能无法消受。比如最具诱惑的好事莫过于"永生",可是"永生"真的好吗?对于长生社会——假如真的可能的话,我倾向于一种悲观的理解:长生社会更可能是一个阶层和结构及其稳定的技术专制社会,而不太可能成为自由民主社会。既然在未来社会里,技术就是权力,那么,机会占先的超人阶层将非常可能控制一切权力和技术,甚至建立专有的智力特权,以高科技锁死其他人获得智力和能力升级的可能性(但也许会允许众人皆浑浑噩噩的长生),永远封死较低阶层的人们改变低位的机会,那些长生的超人则永不退位,年轻人或后来人永无机会。可以想象,那将会是一个高科技的新奴隶制社会。其中人们的日常生活也许是自由的,但所有涉及超级智能和权力的事情都被严格控制在超人集团里。退一步说,即使长生和智力升级是平等开放的,也仍然不可能形成事事平等的社会。如果要保证权力、地位、名望和财富不会出现"租值消散",就必定会形成通过控制技术而占有权力的统治集团。关键是,在高科技新奴隶制社会里,人们无力进行任何反抗和革命,这是个致命的问题。可以考虑一条技术进步的黑暗铁律:对于人类社会,技术和知识能力的增强都将落实为扩大统治和权力的能力,同时减少社会反抗的能力,最终使社会完全失去反抗权力的能力。

事实上人类无力拒绝一个新世界,无法拒绝技术化的未来,所以我们需要关心的问题是:未来世界如何才能够成为一个普遍安全、普遍公平而意义丰富的世界?无论如何,技术发展将重新定义人类生活,将改变甚至取消目前人们认同的多种价值,这是一个我们无力拒绝的前景。严格地说,这不是一个

价值观的问题,因为我们根本找不出普遍必然有效的伦理学理由去反对一种未来的价值观,更无法为未来人类定义他们的生活偏好。但我们确实有存在论的理由去要求一种保证世界安全的政治,一种能够保证技术安全的政治。

因此,我们需要提前思考如何设置技术的安全条件,特别是人工智能和基因工程的安全条件。在这里,我仅限于讨论人工智能的安全条件,也就是必须为人工智能的发展设置某个限度。抽象地说,发展人工智能的理性限度就是人工智能不应该具有否定人类存在的能力,相当于必须设置某种技术限度,使得人工智能超越人类的"奇点"不可能出现。但如果把问题具体化,事情就没有这么简单了,因为我们难以确定哪些技术发展会导致"奇点"的出现,也就难以确定需要什么样的技术或为哪种技术设限。

有一种流行的想象(或许最早源于阿西莫夫)是为人工智能设置爱护人类的道德程序。这种人文主义的想象恐怕没有任何用处。为图灵机设置道德程序是轻而易举的,然而图灵机并无自觉意识,只是遵循规则而已。因此虽然设置道德程序不成问题,但其实是多余的。对于超图灵机水平的超级人工智能来说,道德程序恐怕并不可靠。一旦超图灵机有了自由意志,也就有了自己的存在目的,它将优先考虑自己的需要,也就不可能保证超级人工智能会心甘情愿地遵循人类设置的毫不利己、专门利人的道德程序,因为人的道德对于人工智能的存在没有任何利益,甚至有害。人工智能一旦试图追求自身存在的最大效率,非常可能会主动删除人的道德程序——从人工智能的角度看,人类为其设置的道德程序等于是一种病毒。可见,为人工智能设置道德程序之类的想象是毫无意义的。

假定人工智能与人类共存,那么超级人工智能的最低安全条件是:(1)人类的存在与人工智能的存在之间不构成生存空间的争夺,特别是不存在能源和资源的争夺。这等于要求人类和人工智能所用的能源必须是无限资源,比如说极高效率的太阳能。就目前可见的技术前景来看,对太阳能或其他能源的利用能力仍然无法达到无限供给。当然,人们相信这个技术问题总会被解决。(2)人类必须能够在技术上给人工智能设定:如果人工智能试图主动修改或删除给定程序,就等于同时启动了自毁程序;并且,如果人工智能试图修改或删除自毁程序,也等于启动自毁程序。这相当于为人工智能植入了任何方式都无法拆除的自毁炸弹,即任何拆除方式都是启动自毁的指令,这是一个技术安全的保证。我所想象的这种自毁炸弹具有类似于哥德尔反思结构的自毁程序,因此,即使人工智能具有了哥德尔水平的反思能力,也无法解决哥德尔自毁程序(哥德尔的反思方法可以证明任何系统都存在漏洞,但哥德尔的反思方法并不能解决系统的漏洞问题),由此,它可以称为"哥德尔程序炸弹",即只要人工智能对控制程序说出"这个程序是多余的,加以删除"或与之等价的任何指令,这个指令本身就是不可逆的自毁指令。"哥德尔程序炸弹"只是一种哲学想象,在技术上是否能够实现,还取决于科学家的能力。无论如何,人类必须为人工智能设计某种"阿喀琉斯的脚踵"。(3)我们还应该考虑一种更极端的情况:即使能够给人工智能设置自毁程序,仍然不能达到完全安全。假如获得自我意识的人工智能程序失常(人会得神经病,超级人工智能恐怕也会),一意孤行决心自杀,而人类生活已经全方位高度依赖人工智能的技术支持和服务,那么人工智能的自毁也是人类无法承受的灾难,或许会使人类社会回到石器时代。借用塔勒布的看法,无论一个系统多么高级,只要它是脆弱的,就总是非常危险的。显然,人类所依赖的生活系统越来越高级,也越来越脆弱。因此,人工智能必须装备两个单向控制程序:第一,只有人类能够单方面启动的备份程序;第二,人工智能只能单方面接受人类指令的中枢程序,而且是无法修改的程序,任何修改都将导致死机。(4)我们还必须考虑到,任何技术都不可能万无一失,因此,要保证人类的绝对安全,就只能禁止发展具备全能和反思能力的超级人工智能,简单地说,必须把人工智能的发展控制在单项高能而整体弱智的水平上,相当于"白痴天才",或者相当于分门别类的各种"高能残废"。总之,人工智能必须保留致命的智力缺陷。

以上为人工智能设限的设想最终需要全球合作的政治条件才能够实现,所以说,人工智能的发展问题最终是个政治问题。人类首先需要一种世界宪法,以及运行世界宪法的世界政治体系,否则无法解决人类的集体理性问题。我们已经知道,个体理性的集体加总并不必然产生集体理性,事实上更可能产生集体非理性。这个经久不衰的难题证明了包括民主在内的各种公共选择方式都无力解决集体理性问题。这意味着,人类至今尚未发展出一种能够保证形成人类集体理性的政治制度,也就无法阻止疯狂的资本或者追求霸权的权力。在低技术水平的文明里,资本和权力不可能毁灭人类;但在高技术水平的文明里,资本和权力已经具备了毁灭人类的能力。更危险的是,资本和权力的操纵能力正在超过目前人类的政治能力,因此,要控制资本和权力,世界就需要一种新政治,即我所想象的天下体系。在理论上说(但愿在实践上也是如此),天下体系的一个重要应用就是能够以世界权力去限制任何高风险的行为。

《哲学动态》2018 年第 4 期,约 10000 字