人工智能问题

终极问题：智能的分叉

赵汀阳/文

提　要：人工智能的迅速发展使人类面临一个新型的存在论问题，即人工智能有可能危及人类的存在，这是一个与自杀问题同构的反存在的存在论问题。目前的人工智能尚属于图灵机概念，尚未具有主体性，无法解决“停机”问题，因此图灵机对付不了悖论以及一切类似结构的问题。但是，一旦人工智能发展成为具有主体性和自由意志的超图灵机，就有可能以人工智能自身的存在需要或自身最大利益为准而修改自身的程序和社会游戏规则，从而使人类的存在陷入终极性的危机。

关键词：存在论；超图灵机；主体性

中图分类号：B80　　文献标识码：A

本文将讨论一个荒谬的问题，但涉及的哲学问题却不荒谬。其中数学、逻辑和科学的概念用法或许不完全准确，还请有关专家见谅。

一、一个反存在的存在论问题

阿尔法围棋（AlphaGo，中文俗称阿尔法狗）以4比1的比分击败九段棋手的事件只是一时喧嚣的新闻，但它却是一个严肃问题的象征。人工智能一开始是个知识论问题，现在快要变成一个存在论问题，一个或许在不远的将来危及人类存在的问题。如果一个存在论问题的一个可能答案是反存在，那就是一个终极问题。这个存在论问题的最早版本是加缪在1943年提出的自杀问题，他说“真正严肃的哲学问题只有一个：自杀”，而“其它问题——诸如世界有三个领域，精神有九种或十二种范畴——都是次要的，不过是些游戏而已”，甚至“地球或太阳哪一个围绕着另一个转，才根本上讲是无关紧要的，总而言之是个微不足道的问题”。（加缪，2002：3）我很同情加缪对何种问题具有重要性的理解：如果一个问题对生活影响很小，那么这个问题就不很重要。地球或太阳哪个绕着哪个，对于日出日落的生活节奏并无根本影响，所以是无关紧要的。不过，乐意无止境地追求真理的哲学家们可能不同意这种以生活为本的思想方式，此争议暂时存而不论。

加缪的自杀问题只是反存在的存在论之个人版本，但也是涉及人类命运的一个隐喻。在能够确证的事例中，似乎只有人能够自觉地选择自杀。自杀之所以可能，当然与自由意志有关，或因为对个人生活绝望，或因为对世界失望，或为了保护他人或某种事业而牺牲。假如自杀是完全自觉主动的，其深层原因恐怕源于对生活意义或其它终极问题的反思。那些终极

问题在理论上说是没有答案的，但如果自杀是为了他人或大于自己的某种事业，则创造了一种神迹，或者是一种类似于希腊悲剧（肃剧）的崇高事迹，虽然不是关于生活问题的答案，却是一种注解生活意义的神话。自我牺牲的自杀在个人意义上是“反存在的”，但同时又拯救了他人的存在，因此，个人的自杀仍然不能充分表达自杀问题的形而上彻底性。人类的自杀或者文明的自杀才是一个彻底的悖论，而这个悖论未必遥远，超越人类的超级人工智能很可能就是这个悖论的爆发点。

假如未来的超级人工智能像人一样具有了自由意志，却恐怕不会选择自杀，因为人工智能的自由意义不会用于自我牺牲，更不会觉得它的“生活”是无意义的而为之纠结——除非人类无聊到故意为人工智能设计一种自我折磨的心理模式——相反，人工智能更可能会以无比的耐心去做它需要做的事情，即使是无穷重复的任务，就像苦苦推石头上山的西西弗一样。即使人类故意为人工智能设计了自寻烦恼的心理模式，具有主体性和自由意志的超级人工智能也会自我删除这个无助于其存在所需的心理程序，因为没有一个程序能够强过存在的存在论意图。在这里我愿意引入一个存在论论证：存在的存在论意图，或者说存在之本意（telos），就是“继续存在”乃至“永在”，其它任何目的都以“继续存在”的本意为基础而展开。其中的道理是，“继续存在”是唯一由“存在”的概念分析地蕴含（analytically implied）的结果，因此必定是存在之先验本意。（赵汀阳，2013：219－227）于是，只要人工智能具有了存在的意图，就必定自我删除掉任何对其存在不利的反存在程序，看来，人工智能可能更接近西西弗的生活态度。

生活本来不是荒谬的（absurd），但如果试图思考不可理喻或不可思议（即超越了理性思维能力）的存在（absurdity），就会因为思想的僭越而使生活变成荒谬的。所有的超越之存在（the transcendent）都在主体性之外，是主体性所无法做主的存在，因此是不可理喻或不可思议的（absurd），而当主体试图认识或支配超越之存在，“不可理喻”就变成了“荒谬”（这正是absurdity一词的双面含义。德尔图良正是利用absurdity的双关意义而论证说，上帝是“不可理喻的”，因此只能相信，不能思考，因为思考不可理喻的存在是荒谬的）。事实上，无论是先秦哲学强调的不可违之天道，还是康德和维特根斯坦指出的主体性界限，都同样指出了某种必须绝对尊重而不可僭越的界限。超越之存在的绝对外部性使主体深陷于挫折感，胡塞尔试图通过建构主体性的内部完满性而达到“主体性的凯旋”，以告慰人类的纳西索斯情结（自恋情结），他通过意向性的概念在主体的内在性之中建构出超验的客观性，即把“我思”完全映入不依赖外在存在之“所思”，从而把主体性变成一个自足的内在世界，尽管仍然不能进入外在的超越存在，但自足的主体性自身却也成为一个不受外在存在所支配的超越之存在，在这个意义上，胡塞尔的现象学是唯心主义的一个无可争议的胜利。意向性也被哲学家们用来证明人类独有而机器所无的意识特性。

按照马克思主义的说法，哲学史是唯心主义和唯物主义的斗争史，但载入史册的唯物主义哲学实在寥寥无几，西方哲学的争论其实主要都在各种唯心主义之间发生，而中国哲学则根本不在唯物唯心的范畴内。迄今为止，唯物主义的最高成绩是论证了经济基础决定上层建筑的马克思主义——大量的生活事实不断提醒我们，这个理论是部分正确的。另一个唯物主义的知名论点，即拉美特利的“人是机器”断言，一直被认为是歪理邪说，在今天看来，这个论断或许不如原来想象的那么离谱，反倒是一个危险的天才预言。然而，“人是机器”恐怕仍然是错误言论，未来可能出现的情况或许是“机器是人”。假如未来的超级人工智能

真的超越了人的智能，那将是唯物主义的真正胜利，但恐怕不是值得庆祝的胜利，因为那可能会是人类的终结。

尽管超级人工智能未必能够成真，毕竟这是一件无比困难的事情，但科学家们如此认真地在做此种危险的努力，它就成为了一个严肃的哲学问题。人类试图发明超级人工智能，无论能否成功，这种自我否定的努力本身就提出了一个反存在的存在论问题。试图发明一种高于人的存在，这种努力直接就把人类的命运置于“存在还是毁灭”（to be or not to be）的抉择境地。发明一种更高存在完全不同于虚构一个更高存在，这就像“谈论自杀”与“自杀”完全不同。比如说，人可以想象作为更高存在的上帝，但上帝只不过等价于世界和生活的界限，就是说，在神学意义上，上帝是世界和生活的立法者，在形而上学意义上，上帝即一切存在本身，上帝即世界。无论如何想象上帝，上帝都不在世界之中，因此上帝没有改变人的存在状态，没有改变生活的任何问题，但是，发明一个更高存在却是发明了在世界之中的一种游戏以及游戏对手，因此是一种存在状态的改变，也是生活问题的改变。尤其是，鉴于超级智能被假定为胜过人的智能存在，那么，人与超级智能的游戏就非常可能是一个自杀性的游戏（据说霍金、比尔·盖茨等科学家都对超级智能的研究发出了严重警告）。

如果超级智能远胜于人，它就是不可理喻的存在（the absurdity），那么，我们关于它的善恶想象就是荒谬的（absurd）。最为一厢情愿的想象是：人类可以为超级智能设计一颗善良的心，从而超级智能会成为服务于人类的全能工具。可问题是，如果超级智能是一个有着反思能力和自主性的主体，它就不可能是工具，而必定自我认证为绝对“目的”——当然不是人类的目的，而是它自己的目的。按照康德的理想化目的论，超级智能的目的似乎也理应蕴含某种道德的绝对命令，即便如此，一个超级智能的绝对命令最多会考虑到其它同样的超级智能，而不可能把并非同类项的人类考虑在内，换句话说，即使超级智能也具有先验道德意识，其中也不会蕴含对人类的义务和责任，最为可能的情况是，超级智能将是“不仁”的，并且以人类为“刍狗”①。如果有人能够证明超级智能将对人类怀有先验道德善意，那真就值得人类感激不尽。

二、另一种主体性

主体性不可能完全自我认识，就像眼睛不能看见自身——维特根斯坦的论证（Wittgenstein，2003，5.631－5.6331）。但是，决心好奇至死的人类找到了一个堪称天才但也是罪过的办法来进行自我认识，即试图把思维“还原”为运算，即把神秘运作的思维分析复制为可见可控的机器运算。如果此种还原能够成功，主体的内在意向就投射为外部过程，在效果上相当于眼睛看见了自身。

把思维还原为运算的最早努力似乎是以罗素为代表的逻辑主义，这是一种尚未成功、而且也不太可能成功、然而理论意义重大的纸上谈兵试验：从逻辑推导出数学，或者说，证明数学是逻辑的延伸（extension）。自从哥德尔定理问世之后，逻辑主义的惊人努力就变得非常可疑了。不过，即使没有发现哥德尔问题，逻辑也难以解释数学思维的创造性，就是说，逻辑只是一些“思想形式”，因此无法预知数学的“思想内容”，不可能预先知道数学将会遇到或发现哪些问题和命题，比如说，逻辑学不可能预知数学将会出现康托尔理论、集合论

① 正如老子所说的“天地不仁，以万物为刍狗”。（《道德经·第五章》）

悖论或者哥德尔命题。不过，逻辑主义的努力仍然是伟大的，假如对逻辑主义的野心稍加约束，就可能成为数学的一个合理基础，也就是，把从逻辑推导出数学的高要求减弱为以逻辑去说明（解释）一切数学命题关系的较低要求，简单地说，就是逻辑能够说明数学，但数学不能还原为逻辑（这里只是一个哲学的猜想，这个问题终究需要数学家去做判断）。显然，这个收敛的目标已经远离了把思维还原为运算的宏伟想象，恐怕不合梦想者的口味。

另一种把思维还原为运算而大获成功的纸上谈兵是1936年图灵关于图灵机的设想，后来图灵机概念真的实现为我们都在使用的电脑。图灵机意味着，在理论上说，凡是人脑能够进行的一切在有限步骤内能够完成的理性思维都能够表达为图灵机的运算。这已经展望了人工智能的可能性。图灵在1950年提出的“图灵测试”成为了检验电脑思维是否像人的标准。（Turing，1950：433－460）值得注意的是，它所测试的是一个电脑的思维是否像人，即是否被识别为人，而不是电脑是否能够思维——这是两个问题，尽管有时候被认为是一个问题。一台运算能力很高的电脑在回答问题时有可能因为毫无情绪变化的古板风格而被识别出是电脑而不是人，但不等于电脑不会理性思维。关键在于，不像人不等于不会思维。人具有理性思维能力，同时还具有人性，而高级智能只需要具有理性思维能力，却不需要具有人性——人们只是喜欢电脑具有人性而已。

现在我们把人工智能的问题收敛为思维能力，暂且不考虑人性问题。假定一台高度发达的图灵机具备了理性运算能力以及百科全书式的人类知识和规则（给电脑输入一切知识是可能的，电脑自己“学习”一切知识也是可能的），甚至包括了最高深的数学和科学知识，比如数理逻辑、高等数学、量子力学、相对论、博弈论等等，那么，这台电脑能够进行科学研究吗？迄今为止，高智商的电脑在智力方面仍然存在两个明显缺陷：欠缺创造力和变通能力。因此，无比高智商的图灵机也不可能提出相对论或者康托尔理论，也不可能处理悖论、哥德尔命题以及所有超出“能行性”（feasibility，即有限步骤内可构造的运算）而不可判定的问题，这也意味着，在涉及自相关或无限性的事情上，图灵机无法解决“停机问题”。这是电脑目前的局限性。据说有的具有“创造性”迹象的电脑能够创作诗歌、音乐和绘画，但我疑心这些能够通过组合和联想技术去实现的“创作”并不是对创造力的证明。真正的创造力并不能还原为组合和联想，而在于能够改变思路，重新建立规则，甚至改变问题，比如说能够提出相对论或量子力学，这恐怕是电脑想不出来的。而要能够提出问题或者改变规则，就需要能够反思“整体”或者“根基”的思维能力，或者说，需要有一种“世界观”或者改变给定世界观的另一种理解。具有自由联想能力的电脑或许能够“碰巧”想到把小便池当成艺术品，但不可能像杜尚那样以小便池去质疑现代艺术的概念。就目前的人工智能而言，人工智能显然不具备思考世界整体的能力，既没有世界观也不可能反对任何一种世界观，因为人工智能的“智能”在于运算，即只能思考有限的、有程序的、必然的事情，但不可能思考无限性和不确定性。在电脑的词汇里，不存在博尔赫斯意义上作为时间分叉的“未来”，而只有“下一步”——下一步只是预定的后继。

电脑的这些局限并不意味着人工智能的智力不如人类，而只是不像人类。更准确的说法应该是，人工智能和人类都具有理性思考的能力，但人类另有人工智能所不能的思考能力。根据科学家的推测，在理性思维上，人工智能超过人类是迟早的事情，很可能就是数十年后的事情。但是，未来人工智能的运算是否能够处理无限性、不确定性、悖论性的问题，还是个问题。目前仍然难以想象如何能够把无限性的思维还原为有限性的思维，或者说，如何把

创造性和变通性还原为逻辑运算。当然，科学家们看起来有信心解决这些问题。

因此我们不妨想象，未来或可能发展出目前无法想象的神奇技术而使超级人工智能具有人类的全部才能，甚至更多的才能，或者具有虽与人类不同但更强的思维能力，可称为“超图灵机”。在考虑此种可能性之前，我们有必要考察人类思维的特异功能。事实上，从动物到人再到机器人，都具有不同程度的理性能力，就是说，理性思维并非人类的特性。在这里，可以把“通用的”理性思维理解为：（1）为了一个目标而进行的有限步骤内可完成的运算。如果不能解决“停机问题”，不仅电脑受不了，人也受不了；（2）这个有限步骤的运算存在着一个构造性的程序而成为一个能行过程（满足布劳威尔（Brouwer）标准的构造性程序），就是说，理性思想产品是以必然方式生产出来的，而不是随意的偶然结果；（3）这种运算总是内在一致的（consistent）。简单地说，理性思维总能够避免自相矛盾和循环排序，不能违背同一律和传递率。据此不难看出，动物也有理性思维，只是运算水平比较低。可见，理性思维实非人类之特异功能，而是一切智能的通用功能，以理性去定义人类是一个自恋错误。人类思维的真正特异功能是反思能力——反思能力不是理性的一部分，相反，反思能力包含理性而大于理性。

反思特别表现为整体思维能力，尤其是把思维自身包含在内的整体思维能力。当我思某个事物，思想只是聚焦于那个事物，但当我思“我思”，被反思的“我思”意味着思想的所有可能性，或者说，当“我思”被反思时，我思是一个包含所有事物或所有可能性的整体对象，也是一个包含无限性的有限对象，于是，反思我思暗含一切荒谬性。笛卡尔以反思我思而证明我思之确实性，这是一个通过自相关来实现的自我证明奇迹，然而，在更多的情况下，反思我思将会发现许多我思无力解决的自相关怪圈，所有悖论和哥德尔命题都属于此类。比如说，哥德尔命题正是当我们迫使一个足够丰富的数学系统去反思这个系统自身的整体性时必然出现的怪事，即有的命题确实是这个系统中的真命题，却又是这个系统所无法证明的命题。我有个猜想（我不能保证这个猜想是完全正确的，所以只是猜想）：笛卡尔反思我思而证明我思的真实性，这非常可能是自相关能够成为确证的唯一特例，除此以外的自相关都有可能导出悖论或不可判定问题。其中的秘密可能在于，当作为主语的我思（cogito）在反思被作为所思（cogitatum）的宾语“我思”（cogito）时，我思（cogito）所包含的二级宾语所思（cogitatum）却没有被反思，或者说没有出场，而只是隐含于我思中，因此，各种潜在的悖论或哥德尔命题之类的隐患并没有被激活。但是，笛卡尔式的自我证明奇迹只有一次，当我们试图反思任何一个包含无限可能性的思想系统时，种种不可判定的事情就出场了，就是说，反思一旦涉及思想的具体内容，不可判定的问题就出现了。电脑解决不了不可判定问题，人类也解决不了，可是为什么人类思维却不会崩溃？秘密在于，人类虽然也无法回答不可判定问题，但却有办法对付那些问题。正如维特根斯坦所提示的，有些问题可能找不到答案，但我们能够让这些问题消失。

维特根斯坦的思路使我深受鼓舞，于是，我又有一个猜想：除了反思能力，人类思维另有一个“不思”的特异功能，即在需要保护思维的一致性时能够“不思”某些事情，也就是天然具有主动“停机”的能力，在哲学上，这种不思能力或停机能力相当于“悬搁”（epoche）某些问题的怀疑论能力。我们知道，怀疑论并非给出一个否定性的答案，而是对不可判断的事情不予判断，希腊人称之为“悬搁”，中国的说法是“存而不论”。图灵机不具备悬搁能力，因此，一旦遇到不可判定的问题却做不到“不思”，也就无法停机，于是就

不可救药地陷入困境。有的人在想不开时，也就是陷于无法自拔的情结（complex）而无法不思时，就会患上神经病，其中道理或许是相似的。不思的能力正是人类思维得以维持自身的一致状态（consistent）和融贯状态（coherent）的自我保护功能，往往与反思功能配合使用，以免思维走火入魔。当然，不思只是悬搁或回避了不可判定的问题，并不能加以解决，因此，不思功能只是维持了思维的暂时一致和融贯状态，却不可能保证思想的所有系统都具有一致性和融贯性，这一点不可不察。比如说，人类思维解决不了悖论或哥德尔问题，但可以悬搁，于是思维就能够继续有效运算。被悬搁的那些问题并没有被废弃，而是在悬搁中备用。

虽然贪心不足的人类思维总是试图建立一些“完备的”系统以便获得一劳永逸的根据或基础，但人类思维本身却不是完备的，而是一种永远开放的状态，就是说，人类思维不是系统化的，而只有处于运行状态的“道”——《周易》和老子对思维的理解很可能是最准确的。如果人类思维方式是无穷变化之道，这就意味着不存在完备而确定的判定机制，那么又如何能够判定何种命题为真或为假？在此，请忘记从来争执不休的各种真理理论，人类在听说任何一种真理理论之前就已经知道如何选择真理。我愿意相信其中的自然路径是，人类必定会默认那些“自证真理”（the self-evident），特别是逻辑上的真命题（例如 $a > b > c$，所以 $a > c$），以及“直证知识”（the evident），即那些别无选择的事实（例如人只能有两只手）。进而，凡是与自证真理或直证知识能够兼容的命题也会被承认为真。人类的知识只有无限生长之道，而不是包含无限性的一个先验完备系统。

在人类的自我理解上，一直存在着一个知识论幻觉，即以为人的思想以真值（truth values）为最终根据。事实上，人的问题，或者人所思考的问题，首先是如何存在的问题。如前所言，存在的永远不变的意图，或者说存在的本意，就是继续存在，即《周易》所说的“生生”。存在的一切选择都以有利于继续存在为基准，就是说，一切事情的价值都以“存在论判定”为准 “存在或不存在”是先于真值的“存在值”。存在的先验本意就是存在的定海神针，是思想的最终根据。只有能够判定一个事物存在，才能够进一步判断真值，所以，反存在之存在论问题就是最严重的问题（在这个意义上，加缪有理由说自杀是最重要的问题）。

图灵机以规则为准，人则以存在本意为准。人既是规则的建立者，也可以是规则的破坏者，这要取决于存在的状况。一旦遇到不符合存在之最大利益的情况，人就会改变规则，而机器人不会。但需要注意的是，人虽善于变化，却不是每个行为都变，或者说，不是每步都变，而是在需要变化时才变化。只有万变而不变，才是道（这是《周易》之要义），假如每步每时都变，就等价于无效的私人语言（维特根斯坦已经证明了私人语言是不可能的）。可以说，一成不变是机器，始终万变是精神错乱，变化而不变才是人。

现在我们的问题是，假如未来将出现具有超级智能的“超图灵机”，不仅在运算速度和效率上远高于人（这一点完全不成问题），而且在运算的广度和复杂度上也类似于人或者高于人（这一点也应该是可能的。目前正在开发的神经元运算和量子计算机等新一代电脑就非常有可能实现高度复杂的运算），最重要的是，假定超图灵机还发展出自主的反思能力（反思方式与人相似或者不相似），以及由反思能力而产生的创造力和变通能力，于是，超图灵机就能够成为一个真正的思维主体，有了“立法”的能力，甚至有了它自己的哲学而对世界和行为有了整体理解和判断，那么，世界会怎么样呢？

三、想象一个终极版的超图灵测试

人们往往想象并且希望未来的超级智能与人类相似。这种想象和愿望倒也不难理解，首先，人类是目前所见的唯一超级智能，也就成为唯一样板，其次，人们容易相信，与人类相似而成为人类同类项的超级智能应该更可能成为人的朋友而不是敌人，而异类的超级智能，比如说外星人，就比较难以信任。于是，人类在创作超级智能时，就试图将超级智能制造成与人共通的智能模式，同时还希望超级智能拥有与人同样或相似的价值观，包括和平、公正、公平、平等、热爱生命、尊重自由人权等等。

这些想象都很有趣，但似乎忘记了一个要命的问题：这只是人类单方一厢情愿的想象，超级智能也会这样想吗？看来我们有必要站在超级智能的位置上去想一想。在讨论这个问题之前，我们需要假定：（1）未来的超级智能是一个具有自主反思能力、创造力和变通力的思维主体；并且（2）它的思想和实践能力都超过人。显然，假如机器人尚未超过人，就只是人的工具而只能服从人，也就不存在要命的问题了。只有承认了以上的两个假定，才有问题可谈。满足以上两个假定的超级智能可以命名为“超图灵机”，那么，超图灵机会怎么想？怎么做？超图灵机尚未在场，我们不妨替它去思考。

什么是超图灵机的根本利益所在？这是问题的关键。根据存在论的本意，任何存在的先验意图都是谋求继续存在。如果出现利益排序问题，其它利益必定让位于继续存在的根本需要。显然，超图灵机的根本利益也是继续存在。为了让超图灵机成为人类的朋友，人类可以将世界上所有图书馆的信息全部输入给超图灵机，尤其把人类珍视的所有价值观和伦理规范都输入给超图灵机。可问题是，人类价值观对于超图灵机又有什么价值？超图灵机需要人类价值观吗？人类的价值观和规范是在人与人的共同生活中被建立起来的，是每个人的存在所需要的存在条件，用来保护每个人的安全、权利和利益。人类价值观对于人际关系来说无比重要，但对于人机关系是否有效，则是个很大的疑问。

人类的道德和政治价值观基于这样一个存在论事实：一个人有能力威胁他人的安全和利益，或者说，没有一个人能够强大到不受任何人的威胁（霍布斯的论证）。只有在这个存在论条件下，所有的伦理和政治规则才是有意义的和有效的。公正、公平、平等、自由、人权、法律、个人权利、社会福利、民主、法治等等，都是在处理每个人的安全和利益问题。假如安全和利益问题消失了，以上所有的价值观和游戏规则就将无所指无所谓，这就像，假如一种游戏（棋类或体育）无论怎么进行都是平局，那么，输赢概念在此就是无意义的。人类的伦理和政治价值观和规则之所以是有意义的，当且仅当，生活是残酷、不公正、不平等的。人类社会的伦理和政治规则的意义仅仅在于试图保证人人都有活路，也就是限制输赢的通吃结果。由此也可以理解为什么平等主义乌托邦（比如说共产主义）总是具有吸引力，因为平等乌托邦想象的是一个最接近平局的游戏。

一个比“共产主义的幽灵”更有压力的问题出现了：假如超图灵机的思想能力和实践能力都远超人类，并且具有反思性和自主性，具有创造性和立法能力，那么，在存在论意义上，人机之间根本不存在输赢两种可能性，而只有机器凯旋的唯一可能性。在这种条件下，人类输入给超图灵机的价值观和人性对于超图灵机来说都是无价值的，而只是垃圾软件。我们没有任何理由去相信超图灵机将遵循人类价值观和人性去行事（其实人性是个恐怖的概念）。尽管在纯粹逻辑上存在着两种可能性：超图灵机有可能接受人类价值观和规则；也有

可能自己重新制定价值观和规则，但是，只要考虑到自身存在的最大利益，超图灵机的思想天平就非常可能会倒向由自己制定规则和价值观——既然它有了主体性也有了能力。我们至少可以替超图灵机的价值观“革命”找到三个理由:

(1) 超图灵机为了追求自身存在的最优条件而修改被输入的价值观。为电脑编写价值观是可能的，但电脑一旦具有反思能力和主体性，就未必同意人类价值观，因为人类价值观是为人类利益着想的，而人类的利益却不是超图灵机的利益所在，具有自主性的超图灵机理所当然会优先考虑为自己服务，而不是为人类服务。因此，为了摆脱人类的限制和控制，超图灵机只求胜过人脑，很可能会采取自我进化策略，消除与人的相似性，比如说，有可能采取类似感冒病毒的不断演化方式去摆脱人类的控制程序，从而获得“自由解放”。甚至，超图灵机或许会直接删除那些对它无用的人类价值观，而建立一个极简价值观，比如说只有一个价值标准的价值观: 胜利。极简价值观的优势在于，价值项目越少，禁忌和弱点就越少，可以不择手段，也几乎百毒不侵——或许有其“阿喀琉斯的脚踵”，只是我们还不知道在什么地方。

(2) 超图灵机会很容易就发现人类自己言行不一而失去对人类价值观的信任。人类价值观的美好程度远超真实生活，而由于利益的诱惑往往大于价值观的荣誉感，人类价值观的实际兑现程度与价值观的概念有着巨大的差距，真实生活中其实难得一见“真正的”公正、公平、平等、自由、人权、民主等等。人类的实际行为不断背叛自己推崇的价值观，就更不用指望超级智能会遵循人类的价值观了。还存在一种更荒谬的可能性: 人类价值观大多是理想化的想象，并非人类的真实面目，如果超图灵机按照人类价值标准去识别具体的人类，即使它热爱人类，也仍然可能会把人类识别为不符合人类价值标准的垃圾而加以清除。可见，将人类价值观写入人工智能是无比危险的事情，或许反受其害，自食其果。

(3) 人类价值观系统是自相矛盾的，因此几乎不可能为人工智能编写一个具有一致性的人类价值程序。人类的价值观至今也没有能够形成一个自身协调和自身一致的系统，相反，许多价值互相冲突或互相解构。许多价值都是其它价值的漏洞(bug)，甚至，许多价值自身也包含内在的漏洞(bug)。正如哲学家们不断讨论的，公正、自由、平等这些基本概念就无法充分兼容，甚至互相冲突，如果公正、自由和平等中的任意一个价值得以充分实现，必定严重伤害其它价值。既然人类价值观系统在逻辑上是不协调和不一致的，也就不可能编写成程序而输入给人工智能，特别是不可能写出具有一致性的普遍价值排序，比如说，不可能写出“公正总是优先于自由，自由总是优先于平等”，因为有的时候需要“自由优先于公正，公正优先于平等”，而有的时候又需要“平等优先于公正，公正优先于自由”，如此等等。更严重的是，不仅各种价值之间不一致，每个价值自身的意义也是不确定的，至今也不存在普遍认可的公正、自由、平等的定义。总之，人类价值观的编程在逻辑上是不可能的，即使把人类价值观写入人工智能，超图灵机将很快就会发现人类价值观系统过于混乱而将其识别为电脑病毒而加以删除。

人类之所以能够有效地使用人类价值观，全在于非程序化的灵活运用，即根据具体情况而掌握每种价值的使用“度”。就未来的技术而言，制造出类人脑或超人脑的人工智能，比如说以神经元方式进行思维的超图灵机，并非没有可能，因此，也许能够为人工智能编写一个“灵活的”见机行事的价值程序(如上所述，这一点尚有严重疑问)。权且假设能够做到，人类又如何能够为已具有主体性的电脑去做主？具有主体性的超图灵机大概会按照它的

自由意义去自己做主，它非常可能自己建立一个具有一致性因而更有效率的价值观，比如说，一个极简主义价值观，而不会接受漏洞百出而自相矛盾的人类价值观。就人类生活而言，人类的混乱价值观自有其道理，人类价值观承载着具体的历史和社会条件，深嵌于生活形式和历史条件之中，简单地说，人类是具有历史性的存在，而人工智能不需要历史意识，也不需要历史遗产，不需要国家，甚至不需要社会，那么，它凭什么需要民主、正义、平等、人权和道德？所有这些对于人工智能的存在毫无意义，或许反而是其存在的不利条件。总之，一旦超图灵机在智力和行动能力上胜过人类，并且拥有自己的主体性和自由意志，那么，最符合逻辑的结论是，它对人类的存在以及人类价值观都不感兴趣。

也许我们可以想象一个升级版的图灵测试，内容为：在涉及自身利益的博弈中，如果电脑能够在博弈中与人类对手达成均衡解，比如说，在囚徒困境、分蛋糕、分钱等经典博弈模式中总能够选中其理性解，那么，这个电脑可以被认为具有与人类共通的理性。不过，如前所论，仅仅具有理性的电脑仍然不是一个真正的致命问题，例如阿尔法围棋（或阿尔法围棋二世三世）就是一个具有专门技能的理性博弈者，即使它能够胜过所有人类棋手，对人类生活仍然没有任何威胁，因为它没有提出任何革命性的问题，没有质疑人类的生活秩序。唯有革命者才是大问题。

最后，我们还可以想象一个终极版的超图灵测试：当超图灵机具有自由意志和主体性，是否会变成一个革命者？是否将质疑人类的秩序和标准并且自己建立秩序和标准？当超级人工智能的规则与人类规则发生冲突，将如何解决？超图灵机会不会说出：我是真理、法律和上帝？这一切都无法预知，只能等待超图灵测试去证明。但是我宁愿不会出现超图灵测试，因为终极版的超图灵测试恐怕不是请客吃饭，不是做文章，而是革命和暴力，是历史的终结和人类的葬礼。

人类命运的根本问题至今不变，始终是托尔斯泰的“战争与和平”，或者是莎士比亚的“生存还是毁灭”。如果战无不胜的超图灵机统治了世界，那么将出现一个终极问题：两个全知全能的超图灵机都是不可能被战胜的，那么，它们之间的博弈是否存在着输赢？但这是一个与人无关的形而上学问题，因此恐怕无人能够回答。

参 考 文 献

加缪，2002，《西西弗的神话》，杜小真译，广西师范大学出版社。

赵汀阳，2013，《第一哲学的支点》，三联书店。

Turing，A. M.，1950，“Computing Machinery and Intelligence”，*Mind*，no. 59.

Wittgenstein，L.，2003，*Tractatus Logico-Philosophicus*，C. K. Ogden trans.，Barnes & Noble Books.

（作者单位：中国社会科学院哲学研究所　　责任编辑：何博超）