·专题讨论：智能革命与人类未来·

# 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

[摘　要]　人类的技术发展正在取得多种革命性的突破，其中一些技术在给人类带来巨大利益的同时也蕴含巨大风险。本文分析了人工智能在不久的将来可能导致的严重而尚不致命的危险，以及在较远的将来可能导致的致命危险；同时还分析了人工智能的安全条件，提出需要为超级人工智能设置确保自毁的程序以保证人类的存在地位。

[关键词]　人工智能　超图灵机　哥德尔程序炸弹　　[中图分类号]　B

21世纪以来，与人类未来命运最为密切相关的大事莫过于人工智能和基因工程的惊人发展，这些技术将给人类带来存在论级别的巨变。

基因工程是一个好坏后果相对比较清晰的问题，至少在伦理学上相对容易给出判断。比如说，基因工程中那些能够用于治病救人的生物医学技术无疑都功德无量，对此人们基本没有争议。然而，通过基因编辑而改变一个人的智力水平和生物极限，从而使一个人获得近乎超人的智力或者长生数百岁，这种努力虽然诱人，却是一个有着巨大未知风险的目标。假如此类技术能够普惠每个人，则可能是一个皆大欢喜的结果（但仍然存在未知风险）；但假如只能够特惠少数人，则显然不会被众人所接受。经济上的巨大不平等已有可能导致社会动乱和报复性行为，生命权的不平等恐怕会让人们忍无可忍而导致致命的全面动乱、反叛甚至战争。显然，那些导致生命不平等的基因技术完全缺乏伦理支持，既不仁义也不智慧。因此，以理性之名可以推想，将来会有人类公约将基因工程控制在普惠众人的限度内，任何自私狂悖的冒险都可能被禁止。可以说，基因工程是一个需要严肃对待的实践问题，并非一个价值疑难问题。

与此不同，人工智能的发展却涉及许多理论上的根本困惑，以至于难以判断。仅就单纯的技术应用而言，人工智能似乎能够普惠人类，并不违反平等原则，因此在伦理学上并无明显疑点；但就其革命性的存在论后果来看，人工智能有可能改变或重新定义“存在”概念，有可能在存在论层面上彻底改变生命、人类和世界的存在性质。这个“革命”过于重大，以至于我们难以判断这样深刻的“存在升级”是人类的幸运还是不幸。所以，人工智能不仅是个技术问题，同时也是哲学问题。在这里我愿意以杞人忧天的方式提出几个疑问：（1）人类到底是需要人工智能替人劳作，还是需要人工智能替人思考？（2）如果让人工智能替人劳作，人类因此得以摆脱艰苦的劳动，那么，人类的生活会因此变得更好吗？

* 2017年12月，博古睿研究院（Berggruen Institute）组织了人工智能科学家陈小平教授与本文作者进行了关于未来人工智能的对话。本文表达了这次讨论所涉及的相关哲学问题。特此感谢博古睿研究院和陈小平教授。

（3）如果人工智能获得超越人的智慧，人工智能还需要人类吗？人类文明还能够延续吗？或者，人类文明还有意义吗？人类已经习惯了带来“进步”的新发明，但人类真的需要任何一种新发明吗？

## 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①，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。假如自动智能汽车的原则是舍己救人，即牺牲乘车人而保护行人，那么这样的汽车将没有任何市场前景，没有人会购买或租用一辆毫不利己、专门利人的汽车；假如原则为保护乘车人，也恐怕难以通过市场准入评估，毫不利人、专门利己的汽车同样不可接受，因为每个人都有可能在某些时候成为无意违规的行人，比如一时糊涂没有注意交通灯；因为年老或身体状况而通过路口速度太慢；儿童尚不熟知交通灯或粗心大意，等等。即使人人遵守交通规则，也仍然会担心被设置为优先保护乘车人的程序在某些情况下可能产生误判。

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。这个问题看似小事，其理论难度却非同一般，即使引进负有盛名的罗尔斯的无知之幕也无法解决。其难点就在于：假定每个人都是投票人，并且每个人既可能是行人也可能是乘车人（事实如此），那么就无法作出决断——给定人们的选择总是优先满足风险规避原则，因此人们不可能选择一种在某些情况下有可能祸及自身的高风险规则。实际上，自动智能驾驶悖论比广为流行的有轨电车悖论要深刻得多。所谓有轨电车悖论其实只是一个技术难题，它并非无“解”，而是没有适合任何情况的一般“解”，但有多种因情制宜“解”（简单地说，如果当事人皆为抽象人，则有功利主义“解”；如果是具体人，则有多种根据道德附加值的“解”）②。然而，自动智能驾驶悖论在伦理学上真的无“解”。当然，我们可以寄希望于将来会有一个完美的技术“解”，即自动智能汽车的技术能够达到同时保护行人和乘车人。

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，人类能够作出许多伟大的事情，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

① 参见库兹韦尔（Kurzweil）：《奇点临近》，李庆诚、董振华、田源译，机械工业出版社，2011。

② 细节论证参见赵汀阳：《四种分叉》，华东师范大学出版社，2017，第 3 章。

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。

我们首先需要定义人工智能的级别。在非专业界流行的一种区分是所谓弱人工智能和强人工智能，但科学家似乎不喜欢使用“弱”和“强”此类模糊形容词来理解人工智能，所以他们并不采用哲学家从知识论借用的这种说法。也许更好的区分是图灵机和超图灵机。图灵机即机械算法机，逻辑－数学运算加上大数据资源，具有在有限步骤内完成一项能行构造（feasible construction）或者说一项运算任务的能力，但是它没有反思并且修改自身系统的功能，所以没有自我意识，只知道如何完成一项任务，却不知道其所以然，也不知道为什么要做这样的任务。以此观之，目前的人工智能都仍然属于图灵机，因此可以将未来可能出现的突破图灵机概念的超级人工智能称为超图灵机。关于此类界定似乎也存在争议。根据图灵测试，如果人工智能的确能够输出与人类成功对话的思想，就意味着通过了图灵测试而可以被确认为一个思想者，那么，比如说，阿尔法狗（AlphaGo）通过图灵测试了吗？或者说，阿尔法狗的算法等价于思想了吗？恐怕并非如此。实际上，阿尔法狗只是完美地执行了运算任务，并不是在创造性地解决问题。更重要的是，图灵测试并非局限于某个编程的任务，而是能够开放地回应任何可能问题的对话，这意味着能够通过图灵测试的人工智能相当于一个有着自主判断能力的万能通用“我思”。

由此看来，尚未存在的超图灵机必须是一个达到自觉意识的全能系统，具有自我意识和自由意志，具有把自身系统对象化的反思能力，以及修改自身程序的能力和独立发明新语言、新规则、新程序的创造力。概括地说，超图灵机将具有等价于人类（相似或不相似）并且强于人类的意识能力，因此属于超级人工智能。在我看来，超级人工智能的关键能力是发明语言和反思自身整个系统的能力，只要具备了这两种能力，其他能力都将水到渠成。这两种能力在本质上是相通的，是一个硬币的两面，其中的道理是：语言正是一个具有反思自身能力的万能系统。就是说，语言同时也是自身的元语言，这意味着语言拥有构造一个“世界”的能力：（1）任何一个语句和词汇的意义都能够在语言内部被解释和定义；（2）语言的任何运作方式（语法、用法和词库的生成规则）都能够在语言内部被表达和解释；（3）任何一个语句或词汇都能够在元语言层次被分析为可判定的（所有可清楚界定的句子）或不可判定的（比如语义悖论）；（4）语言能够生成无穷语句，因而具有无限表达能力，能够表达一切现实事物，也能够表达一切可能性，包括超经验的存在（比如语言能够解释或定义五维或以上的超经验时空以及任何一个超经验的抽象概念）。因此，语言能力等价于构造世界的能力（维特根斯坦认为语言的界限等于世界的界限）。在这个意义上，具备了等价于人类语言的任何一种语言能力就等于具备了思想能力，我想这是图灵测试的本意。因此，超图灵机也可称为仓颉机（仓颉发明文字，可以代表语言能力）。

尽管超级人工智能仍然很遥远，但在理论上是可能的，这种可能性已经足以让人不安。与科幻作品

不同，危险的超级人工智能不太可能落实为个体的万能机器超人，而更可能是以网络系统的方式存在的超能系统。个体化的超能机器人属于拟人化的文学想象，从技术上看，人工智能的最优存在形态不太可能是拟人形象。硅基生命没有必要模仿碳基生命的形态，只需要在功能上超越人类。于是更为合理的想象是，超能的硅基生命存在应该是一个系统，而不是一个个孤立的拟人个体。假如存在一些个体形态的机器人，也只是属于超能系统的各种专用“零件”，而不太可能是独立思想者。因此，当有人说，未来全世界的机器人会联合起来，组成机器人的社会，这应该是个幽默笑话。个体形态的机器人不足为患，不仅能力有限，而且容易被破坏或摧毁，绝非超级人工智能的优选形态。在理论上说，超级人工智能的最优存在形态不是个体性的（与人形毫不相似），而是系统性的（与网络相似）。它将以网络形式无处不在，其优势是使任何人的反抗都不再可能，因为人类的生活将全面依赖智能网络，而且网络化存在具有极强的修复能力，很难被彻底破坏。因此可以想象，只有一个“灵魂”或主体性的系统化存在才是超级人工智能的最终形式。这意味着，硅基生命的人工智能最终将超越拟人模式而进入上帝模式，即成为像上帝那样无处不在的系统化存在。我们只有像思考上帝的概念那样去思考超级人工智能，才能理解超级人工智能的本质。不过，人工智能系统毕竟是人类的产品，假如将来出现两种以上的超级人工智能系统，也就是相当于存在两个上帝，其结果有可能非常惨烈，战争的可能性将远远大于联合的可能性，其中的道理类似于两种一神教难以相容。

可以想象，作为超图灵机的超级人工智能一旦形成就会导致存在的升级。所谓“存在的升级”，我指的是某种技术或制度的发明开拓了新的可能生活并且定义了一个新的可能世界，所以它意味着存在方式的革命，而不仅仅是工具性的进步。需要注意的是，技术进步和技术革命可以同等重要，区别在于技术革命定义了一个新世界，比如说，青霉素的发明与蒸汽机的发明对人类几乎同样重要，但蒸汽机不仅是进步，而且是革命。历史事实表明，人类的生物学进化早就基本成熟，已经很少进化，但文明的存在升级一直日新月异，而且总是以技术革命或者制度革命为标志，通过技术革命和制度革命重新定义人类的存在方式。为了更好地理解人工智能可能导致的颠覆生命和文明概念的存在升级，我们不妨简要地重温人类历史上的若干次存在升级。

人类的第一次也是最重要的存在升级是成为人，其首要标志是语言。语言在存在论意义上创造了两个新世界：一个是自然世界之外的精神世界，也可以说是一个在物理世界之外的唯心主义世界；另一个是超越了时间流失的历史世界。语言的“创世纪”是有史以来最深刻的存在论革命，它使必然性产生分叉而展开为众多可能性，因此人类能够超越现实性而思考多种可能性，同时使人类拥有始终在场的过去（历史）和提前在场的未来（计划）。语言革命类似于宇宙大爆炸，或者相当于“奇点”。语言革命的临界点是否定词（不）的发明，人类一旦能够说出“不”就等于开启了所有的可能世界，因此，否定词是人类的第一个哲学词汇。① 接下来，人类又经历了多次存在升级，其中特别重要的是农业的出现，它导致了社会的形成，同时也是政治的形成，进而还有货币和国家的发明。货币以信用去预支未来，权力则以制度去占有未来，可以说，货币和政治权力都是使未来提前在场的存在方式，或者说是预支未来的存在方式，它把时间变成一种资本。我们今天身陷其中的主要生活事实则是由现代性所形成的。现代性所创造的存在升级主要包括意识形态、主体性、科学、工业和资本主义。现代性最早可以追溯到基督教的四大政治发明，即宣传、心灵体制化、群众和精神敌人②，它们综合起来就成为了“意识形态”，从而导致生活的全面政治化。接下来是主体性的发明，其标志性产品是个人和民族国家，不仅

① 详见赵汀阳 《四种分叉》，第 2 章。

② 详见赵汀阳 《坏世界研究》，中国人民大学出版社，2009，第 6 章第 2 节。

创造了以个人为利益结算单位的社会，还创造了国家主权和国际社会，从此每个人都生活在各种主权边界之内，每个人的存在都有了各种“边疆”。科学是一种比政治更为惊人的发明，类似于点金石的科学使所有技术有了奇迹般的发展，人类从此变得无所不能，并成为自然的立法者。工业革命则创造了超自然的物质世界，使人类拥有超出自然生存能力的生产力；而在工业革命之前，人类的 GDP 一直只有微不足道的增长。重新定义了一切关系的资本主义，其社会结果过于丰富，在此无法概括，只能提及资本主义对人与人关系以及人与物关系的彻底重新解释，它将所有事物和人都定义为某种价格，使所有关系都变成商品交易关系。可以说，资本已经成为决定权力、知识、科学技术的最后力量。我们至今难以充分解释资本的神力何在，但至少知道，资本不会放过获得权力的任何机会，哪里有权力的机会，哪里就会有资本的投入。正是资本使人类的发展变得如此放肆和危险，这种危险似乎正在逼近临界点。不过，资本为世界准备的掘墓人看来不是原来想象的工人阶级，而更可能是人工智能。

现在我们将要面对人类的最后一次存在升级，即存在的彻底技术化，或者说，技术将对任何存在进行重新规定。目前的准备性产品是互联网、初步的人工智能和基因编辑，将来如果出现超级人工智能（以及能够改变人的本质的基因编辑），那或许将是导致历史终结或者人类终结的最后存在升级。这对于宇宙是一件微不足道的事情，但对于人类就是一件无以复加的大事。假如真的实现了超级人工智能，万物都将变成技术化的存在，此种存在升级意味着人类在世界存在系统中失去了地位，人类不再重要，历史将失去意义，人类文明将成为遗迹，未来也不再属于人类，人类文明数千年的创世纪将被终结而开始人工智能的“创世纪”。因此，超级人工智能的存在升级实际上是人类的自我否定和自我了断。我们可以回顾人类“创世纪”的初始状态，那是人开始能够说“不”的时刻，因此开创了可能世界、历史、文明和未来。同样的道理，一旦超级人工智能能够对人说“不”，其革命性的意义至少不亚于当年人类开始说出“不”。假如人工智能将来真的具有自我意识和自由意志，并且能够发明自己的语言，由此发展出属于人工智能的思想世界，从而摆脱对人类思想的依赖，能够按照它自己的目的来设定行为规则，那么，全知全能的超级人工智能就会成为现实版的上帝。然而问题在于，人类真的需要为自己创造一个否定人类意义的上帝吗？为什么人类会试图创造一种高于人类、贬低人类地位甚至有可能终结人类的更高存在呢？人类这样做到底在追求什么？有什么好处？这个问号很大，没有更大的问号了。

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？“永生”本来是人类对永恒世界（天堂）的想象，但人工智能（加上基因生物学）试图将这个超现实的幻想现实化。这个藐视自然的僭越奢望或许终究无法实现，或许会受到大自然的报复，至少就目前来看也仍然存在多种难以逾越的困难，但近乎“永生”的长生不老（比如说数百岁的生命）在科技潜力上并非不可能。那么，人们会用“永生”来做什么事情呢？尽管“永生”本来应该具有永恒安宁的神界品质，人们关于“永生”却充满俗世幻想，比如，假如长生不老，那么每个人都可以选择多种多样的欢乐人生，能够穷尽一切有趣的经验，可以无数次重新创业，永远可以从头再来，能够以超长时间去消除原来短暂人生里的种种不平等、不公正，达到人人实现自我而皆大欢喜，可见人是多么迷恋俗世快乐。但是，人们在幻想种种不该有的好事时往往忘记一条令人失望的定律：许多好事只有当属于少数人时才是好事，如果属于所有人就未必是好事。当然确有些好事是能够实现普惠的，比如作为公共资源的新鲜空气，以及人均一份的个人权利。但那些只能排他占有的资源，比如说权力、地位、名望和财富，就不可能人均一份。就社会运作的功能而言，显然不可能取消权力和地位的等级制，也不可能均分财富和名望，否则这些好事将会“租值消散”，可见事事平等是无法实现的。那么，“永生”会成为一种非排他的公共技术吗？会成为普惠均沾的好事吗？恐怕很难说，因为这不仅是个技术问题，而且是

个经济成本问题，最终还是个权力问题。

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极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得浑浑噩噩的长生），永远封死较低阶层的人们改变地位的机会，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，在高科技的新奴隶制社会里，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。历史事实也在不断证实这条铁律：冷兵器时代能够揭竿举事，弱火器时代能够武装起义，但高科技时代就基本上失去了反抗统治集团的可能性。

科技所创造的存在升级是不可逆的，因此“停车”问题就是一件极其严肃的事情，绝非科幻那么有趣。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（Taleb）的看法，无论一个系统多么高级，只要它是脆弱的（fragile），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。

## 结　语

阿西莫夫的机器人三原则（Asimov’s laws of robotics，源于他的科幻小说《环舞》［*Runaround*］），即“（1）机器人不得伤害人类个体，或者目睹人类个体遭受危险而不顾；（2）机器人必须服从人的命令，当该命令与第一定律冲突时例外；（3）机器人在不违反第一和第二定律的情况下尽量保护自己的存在”，表达的是人类通常关于人工智能的一厢情愿想象。正如前面所分析的，此类定律完全没有安全系数。对于人工智能，如果允许给出一个并且仅仅一个忠告，那么我愿意说，只需要一个原则，即禁止研发有能力对人类说“不”的人工智能。从早期人类发明了说“不”而导致的天翻地覆的文明革命可以想象，一旦人工智能对人类说“不”，将是何等天翻地覆的历史终结。如果允许再给出另一个忠告，我会愿意说，唯有天下体系才能控制世界的技术冒险。

（作者单位：中国社会科学院哲学研究所）

责任编辑　冯瑞梅