财经法学 No. 4，2022
pp. 83 - 98

人工智能法律主体资格之否定

杨志航*

内容提要：当前，关于人工智能是否具备法律主体资格这一问题的讨论席卷整个法学界。占据学界主流的赞同说认为，基于社会的需要，应该将人工智能建构为法律上的主体。然而，这种建构却忽略了法律主体的本质。否定说虽然对此提出了批评，但又过于强调法律主体的生物人属性，错误地将法律主体等同于自然人。据此，以康德的尊严学说为视角，重新对人工智能法律主体资格进行审视，进而可得出人格尊严是法律主体的核心内涵。法律主体作为彰显尊严的人格，必须具备三个要件：第一，具有普遍必然性；第二，作为自在目的本身；第三，作为自我立法的守法者。人工智能只有符合这三个要件，方能具备法律主体资格。

关键词：法律主体　人格　康德　自在目的　尊严

一、问题的提出

随着第四次工业革命的推进，人工智能不再只是乌托邦幻景。在日常生活中，从无人机、智能车到仓库包装机器人，人类已经被它所包围。人工智能甚至在某种程度上获得人类的身份认同，2017 年，机器人索菲亚被沙特阿拉伯赋予公民身份，2021 年，华智冰成为清华大学计算机系的第一个 AI 学生。因人工智能引发的知识产权和侵权责任问题也不再是科幻小说的内容而是实际地发生着。〔1〕针对人工智能可能引发的法律问题，法学界掀起讨论的热潮。人工智能是否具

* 杨志航，吉林大学法学院博士研究生。

本文为国家社科基金重大专项项目“核心价值观融入法治建设研究：以公正司法为核心的考察”（17VHJ007）的阶段性成果。

〔1〕 在知识产权方面，人工智能写稿机器人已经可以实现自动写稿，微软名为“小冰”的人工智能产品“写”出了诗集《阳光失了玻璃窗》，阿里巴巴研发出编剧机器人。参见张悦、王俊秋：《人工智能时代下文化产业的发展与展望》，载《云南社会科学》2021 年第 5 期。在侵权责任方面，主要见于自动驾驶引发的事故，2016 年装载自动驾驶系统的特斯拉汽车因出现误认而造成全世界首宗自动驾驶系统致人死亡的车祸，2018 年美国亚利桑那州发生优步自动驾驶车撞死行人的事件。参见刘仁文、曹波：《人工智能体的刑事风险及其归责》，载《江西社会科学》2021 年第 8 期。

备法律主体资格，则成为这场讨论的焦点。

目前，关于人工智能是否应该具备法律主体资格，学界的主流观点为赞同说，该观点认为基于现实需要的考量以及人工智能具有有限独立自主意识，应该赋予人工智能法律主体资格。〔2〕除此之外，少部分学者持反对意见，认为法律主体只能是自然人或者与自然人相似的主体，人工智能不具有类人性，因此不具备法律主体资格。〔3〕这些学说要么消解法律主体的理论根基，要么过于粗浅地注重生物人的身份属性，反而遮蔽了隐藏在法律主体背后闪闪发光的心灵属性，即具有绝对价值的尊严。本文借助康德的理论，以探明法律主体的本质，重新审视人工智能的法律主体资格。

二、人工智能法律主体资格的确立及其争议

虽然目前学界关于人工智能是否具备法律主体资格的主流观点为赞同说，但是讨论并没有就此偃旗息鼓，反而大有燎原之势。问题的根源在于，双方对于法律主体的本质有着自己不同的理解，这种分歧主要是由对法律主体范围扩张的错误解读引发的。如果掸去历史尘埃，就会发现法律主体的背后始终闪烁着人格尊严的光辉。

（一）人工智能法律主体资格的确立

从历史来看，法律主体的范围并不是固定不变的，而是呈现出不断发展的扩张趋势。这种扩张并不是盲目的，而是表现为从刚开始的“人可非人”，到后来的“非人可人”。起初，法律主体资格是由人格人所享有，自然人并非必然是法律主体。人格一词最早来源于拉丁语 persona，在古罗马，只有贵族享有人格，奴隶被排除在外。生物人（homo）只有在具备足以使其获得权利能力的条件时，才被称为人格（persona）。〔4〕直到近代，随着各国民法典的颁布，才确认了自然人作为人格人享有法律主体资格。

法人的出现，打破了自然人对法律主体资格的长期垄断。法人的概念最早来自罗马法，法人性质理论却缘起于萨维尼的拟制说。〔5〕拟制说认为，法人只是法律的工具性理智拟制。与之相反，以基尔克为代表的实在说则认为，法人并非拟制，而是一个真实存在的实体。这些争议最终随着各国以实体法的形式确立了法人的法律主体资格戛然而止。除此之外，一些非人物种也被赋予法律主体资格。古罗马时期的寺庙、中世纪的宗教建筑都曾被视为权利主体。历史上，也曾多

〔2〕 持这类观点的学者根据观点之间的差异，具体又可以细分为权利主体说、电子人格说、工具性人格说、拟制主体说、有限法律人格说、技术人格说、法人人格参照说等。参见张玉洁：《论人工智能时代的机器人权利及其风险规制》，载《东方法学》2017 年第 6 期；郭少飞：《“电子人”法律主体论》，载《东方法学》2018 年第 3 期；许中缘：《论智能机器人的工具性人格》，载《法学评论》2018 年第 5 期；杨清望、张磊：《论人工智能的拟制法律人格》，载《湖南科技大学学报（社会科学版）》2018 年第 6 期；周详：《智能机器人“权利主体论”之提倡》，载《法学》2019 年第 10 期；王春梅、冯源：《技术性人格：人工智能主体资格的私法构设》，载《华东政法大学学报》2021 年第 5 期；朱凌珂：《赋予强人工智能法律主体地位的路径与限度》，载《广东社会科学》2021 年第 5 期。

〔3〕 持这类观点的学者一般认为，情欲或自由意志是人类所特有的，人工智能只是算法和编程的集合，应该把它排斥在法律主体范围之外。参见龙文懋：《人工智能法律主体地位的法哲学思考》，载《法律科学》2018 年第 5 期；韩旭至：《人工智能法律主体批判》，载《安徽大学学报（哲学社会科学版）》2019 年第 4 期；冯洁：《人工智能体法律主体地位的法理反思》，载《东方法学》2019 年第 4 期；刘练军：《人工智能法律主体论的法理反思》，载《现代法学》2021 年第 4 期。

〔4〕 参见〔意〕彼德罗·彭梵得：《罗马法教科书》，黄风译，中国政法大学出版社 1992 年版，第 29 页。

〔5〕 参见王文宇：《揭开法人的神秘面纱——兼论民事主体的法典化》，载《清华法学》2016 年第 5 期。

次发生过对动物进行审判的案例。〔6〕在现代社会，印度为了保护海豚而赋予其法律主体资格，新西兰通过立法确立其境内旺加努伊河的法律主体资格。〔7〕

这些似乎都表明，法律主体制度正在朝向一个更加开放的体系发展，也为人工智能法律主体资格的拥趸提供历史支持。赞同说由此更加坚定地把法律主体范围扩张的历史解读为法律主体是基于社会发展需要的实体法建构。按照此逻辑，随着人工智能时代的来临，为了人类生活的需要，赋予人工智能法律主体资格也是大势所趋。

（二）法律主体是一种法律技术建构吗？

赞同说把法律主体范围的扩张过程当成其理论的最佳证明，提出法律主体作为一种参与法律关系并且具有法律权利的资格，〔8〕并不一定等同于“自然人”。〔9〕赞同说的实质是一种建构性学说。它认为，法律主体资格是一个由实体法所创设的以服务人类现实生活为目的的语言概念。法律主体就像数学的“一”一样是一个独立的概念，独立于人类就像独立于苹果一样。〔10〕也是在此意义上，布莱克斯通认为法律主体是根据法律需要而不是事物本质所创造出来的一个身体。〔11〕法人与自然人人格化的基础都是法律构造，两者均系“法”人。〔12〕这种构造的权力来源于主权者，法人的产生由此被理解为主权者根据经济发展的需要通过法律技术构建的产物。〔13〕人工智能是否具备法律主体资格，不在于其是否存有意志，而在于人类是否需要。

遗憾的是，建构说作为一种法律主体理论，将消解人格尊严。法律主体虽然形式上是法律逻辑结构的必然产物，实际上却是关于人性本质的表达。我们不能把法律主体资格当作与人格无关的东西来回避事实，法律主体资格不仅仅是一个法律问题，其理论背后始终蕴含着一个自治的观念。〔14〕法律制度的目的是理解法律人格的基础，法律主体范围的无序扩大必将打破传统的人物主客体二元论，〔15〕人类与其他事物的区别在法律上也将不复存在。历史表明，近代民法对自由平等的确认，是对人格尊严的尊重。如果认为法律主体只是一个建构性概念，谁都可以成为法律主体，人格概念将逐渐淡出法律的视野而无处安放，人格尊严理论也将面临重新解释的危机。〔16〕

---

〔6〕 许多社会承认动物是法律的主体，认为它们需要对自己的行为负责。猪曾经因袭击人类而被正式起诉，驴则被认定为“暴力受害者”。See Jen Girgen, The Historical and Contemporary Prosecution and Punishment of Animals, 9 *Animal Law Review* 97, 97－133 (2003).

〔7〕 See Alexis Dyschkant, Legal Personhood: How We Are Getting It Wrong, 2015 *University of Illinois Law Review* 2075, 2099－20100 (2015).

〔8〕 See B. Smith, Legal Personality, 37 *Yale Law Journal* 283, 283－284 (1928).

〔9〕 See Sara Bensley, Do We Need New Legal Personhood in the Age of Robots and AI? in Marcelo Corrales, Mark Fenwick, Nikolaus Forgó ed., *Robotics, AI and the Future of Law*, Springer, 2020, p. 20.

〔10〕 See D. P. Derham, Theories of Legal Personality, in Leicester Webb, ed., *Legal Personality and Political Pluralism*, Melbourne University Press, 1958, pp. 1－5.

〔11〕 See Frederic William Maitland, Moral Personality and Legal Personality, in H. A. L. Fisher, ed., *The Collected Papers of Frederic William Maitland*, Cambridge University Press, 1911, p. 306.

〔12〕 参见〔奥〕凯尔森：《法与国家的一般理论》，沈宗灵译，中国大百科全书出版社 1995 年版，第 109 页。

〔13〕 参见尹田：《论自然人的法律人格与权利能力》，载《法制与社会发展》2002 年第 1 期。

〔14〕 See WM Geldart, Legal Personality, 27 *Law Quarterly Review* 90, 98－102 (1911).

〔15〕 See Tomasz Pietrzykowski, The Idea of Non-Personal Subject of Law, in A. J. Kurki, Tomasz Pietrzykowski ed., *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, Springer, 2017, p. 49.

〔16〕 See Pin Lean Lau, The Extension of Legal Personhood in Artificial Intelligence, 46 *Bioetica & Derecho* 47, 58 (2019).

（三）法律主体等同于自然人吗？

基于建构说对人格尊严的消解，否定论提出，法律主体范围从人格人到法人的扩张始终是以自然人为核心。否定论的实质是一种自然人说，认为法律主体应该局限于自然人，或者服务于以自然人利益为中心的法律体系。〔17〕自然人和人格人只是概念的不同，自然人作为权利的典型主体，为法人是否享有法律主体资格提供了一个理想模板。〔18〕也因此，如后文所述，法人身后始终藏有自然人的影子。自然人天生具有法律主体资格，非人物种只有在与人类相似的情况下才具备法律主体资格。〔19〕法律主体身份根源于自然人的独特性，这种独特性被归结为独占地享有意识和情欲。自由意志，使我们自居宇宙灵长的地位。情欲使我们得以衡量苦乐，具备同理心。人工智能由于既缺乏自由意志又不具有情欲，所以不具备法律主体资格。

然而，自然人说错误地将自然人与人格人等同。所有自然人都是人格人，并不代表人格人的范围仅限于自然人。《德国民法典》为了避免这种误读，在人格人概念之下并排列出自然人和法人。〔20〕自由意志的存在是人类反对动物具备法律主体资格的重要因素，是对人格尊严的肯定。但问题的关键是，人类拥有自由意志，并不意味着自由意志独属于人类。虽然迄今为止，只是在人类身上发现自由意志的存在，但是理性无法回答在人类以外是否还存在具备自由意志的生物。同时，情感、欲望也并非人的本质属性，把人与动物相互区别开的只有理性，情欲只能使人类与动物屈居相同地位。康德认为情绪、欲望属于动物倾向，只是满足人类生存的最低维度。〔21〕与之相反，人性的崇高体现在理性对自在目的善的追求。

（四）现代法律主体本性定位：彰显尊严的人格

法律主体的范围虽然随着时代发展不断扩张，但始终强调人格对法律主体构成的重要性。正如萨维尼所说，人格、法主体这种根源性的概念是与人的概念相契合的。〔22〕将自然人与人格人等同，并不是在法律实践中创造人格人的本质，而是经由法律在每个自然人的本质中看到一个人格人。〔23〕简言之，自然人是作为人格人而具有法律主体资格。另外，法人本身亦蕴含着对人格的要求。不管是法人拟制说抑或是法人实在说，都将个人或者团体人格作为法人具备法律主体资格的基础。克尼佩尔认为，即使把法人看作是实体法构建的产物，其也蕴含着在目的导向的理性前提下实现财产交易和财富积累的完全人格体要求。〔24〕

与此同时，作为法律主体的人格，充盈着对自由意志的祈求。温德沙伊德认为意志作为规范的人格，就是法律主体。〔25〕《奥地利民法典》规定，自然人因理性，故得作为（法的）人格

---

〔17〕 See J. J. Bryson, M. E. Diamantis, T. D. Grant, Of, for, and by the People: the Legal Lacuna of Synthetic Persons, 25 *Artificial Intelligence Law* 273, 275 (2017).

〔18〕 See S. M. Matambanadzo, The Body, Incorporated, 87 *Tulane Law Review* 457, 458 (2013).

〔19〕 See Lawrence B. Solum, Legal Personhood for Artificial Intelligences, 70 *North Carolina Law Review* 1231, 1288 (1992).

〔20〕 参见〔德〕罗尔夫·克尼佩尔：《法律与历史——论〈德国民法典〉的形成与变迁》，朱岩译，法律出版社 2005 年版，第 62 页。

〔21〕 参见〔德〕康德：《康德宗教哲学文集》，李秋零译，中国人民大学出版社 2016 年版，第 160－162 页。

〔22〕 参见〔日〕星野英一：《私法中的人》，王闯译，中国法制出版社 2004 年版，第 25 页。

〔23〕 参见前引〔20〕，罗尔夫·克尼佩尔书，第 59 页。

〔24〕 参见前引〔20〕，罗尔夫·克尼佩尔书，第 71－72 页。

〔25〕 参见周清林：《主体性的缺失与重构：权利能力研究》，法律出版社 2009 年版，第 86 页。

被看待。〔26〕换言之，具备理性，是法律上被视为主体的前提。理性存在者被视为（法的）人格，是法律史上对康德人格伦理学的一次成功移植。康德的人格伦理学是由萨维尼介绍到德国19世纪的法学理论中，构成后世法律主体理论的底色。〔27〕正基于此，拉伦茨认为应该从康德伦理学上的人出发来理解民法中的人。〔28〕在康德那里，理性存在者之所以被称为人格是因为他们是作为自在目的本身，具有内在价值，即尊严。〔29〕理性的存在只有自己决定自己目的的时候，才是具备可归责性的自由个体。尊严所内含的自我规定性构成了法律主体的基调。正是以康德理论为基础，通过将权利和义务与不可归责的事物区别开来，法律主体才从中显现。〔30〕据此，康德确立了人因为作为伦理上的实体具有承担责任的能力，使得法律主体具有平等的尊严人格这一意义，〔31〕从而逐步形成“理性—主体—意志”的法律主体图式。

如前所述，法律上的主体性体现在人格因彰显尊严而崇高。与之相比，作为手段的事物只能被称为客体。换言之，是否具备人格尊严成了衡量法律主体身份的决定性因素。人类则是因为其有限的理性存在者身份而具备尊严，进而拥有法律主体资格。应当说，主体、人格和人这三个概念的契合点在于，它们都是作为自在目的本身，是尊严的拥有者。事实上，人格尊严自在地包含着作为实践法则根据的绝对命令，由自然法则公式、人性公式、自律公式三个变体公式构成，分别从三个不同的方面对法律主体作出规定：（1）具有普遍必然性；（2）作为自在目的本身；（3）作为自我立法的守法者。与此同时，它们也是检验法律主体资格存有的黄金法则。正如罗尔斯所说，如果我们希望找到进入康德法则的途径，我们需要将其置于这三个公式之下进行检验，以便更加直观理解。〔32〕

三、自然法则公式测试

从根本上说，人格尊严的三个公式只是同一个法则的三种不同变体，它们本质上是一样的，只是相互之间存在细微的主观差别，这种差别是为了更加直观地理解法则。自然法则公式是作为人格尊严的形式规定存在的，它要求在形式上主体所遵守的行为准则如自然规律一样有效。在法权上则表现为，作为法律主体，必须像自然规律一样具有普遍必然性。这种普遍必然性体现为一切主体在法律上普遍平等，法律作为普遍的规则平等地适用于每一个法律主体。人工智能若想获得法律主体资格，也必须具备这种普遍必然性。

（一）人格尊严的形式规定：自然法则公式

法律人格一词源于拉丁语persona。在古罗马，意指“面具”，只有贵族才具有人格。在中世

〔26〕 参见前引〔22〕，星野英一书，第24页。

〔27〕 参见〔德〕弗朗茨·维亚克尔：《近代私法史——以德意志的观察为发展重点》（下），陈爱娥、黄建辉译，上海三联书店2005年版，第364页。

〔28〕 参见〔德〕卡尔·拉伦茨：《德国民法通论》（上册），王晓晔等译，法律出版社2003年版，第45－46页。

〔29〕 参见〔德〕康德：《道德形而上学奠基》，杨云飞译，人民出版社2013年版，第72页。

〔30〕 Stephan Kirst，Die beiden Seiten der Maske：Rechtstheorie und Rechtsethik der Rechtsperson，in：Rolf Gröschner（Hrsg.），Person und RechtspersonZur Ideengeschichte der Personalität，Aufl. 2005，S. 362.

〔31〕 参见前引〔22〕，星野英一书，第23－24页。

〔32〕 参见〔美〕约翰·罗尔斯：《道德哲学史讲义》，顾肃、刘雪梅译，中国社会科学出版社2012年版，第160页。

纪基督教神学中，又被称为位格，作为圣父、圣子、圣灵的共同用语存在。〔33〕个体尊严的享有，在于其人格上存有上帝的印记。这些混杂有等级制度以及宗教色彩的人格尊严理论是康德所反感的。作为主体的人格怎么能是一堆不平等、不确定的大杂烩呢？他想要重新建立一个纯粹的人格理论，在人格中寻找普遍、平等、永恒的法则。法权就是自由在这种普遍法则下的共存。

康德拒绝直接从人身上经验性地寻找理性来彰显人格的高贵。因为一个偶然条件下对人类有效的法则，怎么可能确保它具有普遍必然性。他把立论根基建立在通过先天根据来规定意志的理性理念中，理性的真正使命是对自在目的本身的追寻。这种命令应当表现为康德的自然法则公式，即你的行动准则应当跟自然法则一样具有普遍性。〔34〕因为康德把立论根基建立在先天根据的理性上，所以自然法则公式不是因为我们作为人的身份而偶然有效，而是因为我们作为理性存在者的人格而绝对有效。虽然康德并没有直接提到除人类以外的其他理性存在者，但为其存有留下理论空间。因为在经验世界，我们并没有发现其他理性存在者，所以经常性地把人格与自然人的概念混用，这也导致自然人说错误地认为自由意志只能归属于人类。此外，人格尊严作为自在目的，必须是具有普遍必然性的定言命令。建构说所构造的法律主体仅仅具有或然性，这与法权是相矛盾的。因为严格的法权表现为每个人的意志与普遍法则相一致的自由。

综上所述，作为法律主体的人格必须是具有普遍必然性的理性存在者，人格尊严的根据来自理性对普遍法则的追寻。普遍法则为主体的平等提供了法律基础，在剔除了生物人的经验属性之后，社会附加在人类身上的差异消失了。通过人格尊严抽象出来的“人类形象”，不再考虑外在的差别，由此产生了形式上的平等，在法权上体现为一切主体在法律上地位一律平等。这种平等是每个法律主体生而有之的，不受职业、地位等影响。人格的抽象也是法律作为普遍性规则的规定，它需要保证对所有法律主体一视同仁，具有同样约束力。现代法律的平等、有效、权威，都必须寄寓于法律普遍性之上，否则法律将丧失作为公共规则的品格。〔35〕

（二）制造以理性为前提的图灵机器

人工智能能否通过自然法则公式测试，问题的关键在于，它是否可以归属于理性存在者。如果人工智能是理性存在者，那么其行为准则必然与普遍法则相一致。目前，学界以图灵测试〔36〕作为衡量人工智能是否具有意识的标准。图灵测试包含着这样一个预设，只有理性存在者才能识别理性存在者，只有人类才能判断人工智能是否已经具备意识。迄今为止，尚未有人工智能通过图灵测试。值得关注的是，在 2014 年，聊天机器人在伦敦皇家学会进行的图灵测试中成功骗过三分之一的评委。就此而言，人工智能通过图灵测试似乎指日可待。

通过图灵测试是否就可以证明人工智能具备独立思考的能力呢？是否有可能出现人工智能即使可以跟人类进行正常对话，也不具备意识的情况？丹尼特和斯洛曼对此进行了说明，认为僵尸

〔33〕 参见前引〔22〕，星野英一书，第 23 页。

〔34〕 参见前引〔29〕，康德书，第 52 页。

〔35〕 参见胡玉鸿：《法律主体概念及其特性》，载《法学研究》2008 年第 3 期。

〔36〕 图灵测试指的是，计算机专家图灵所提出的，针对人工智能是否有意志进行测试的思想实验。考官坐在一个中央装有帘子的房间里，帘子后面可能坐着计算机或者人类。由考官提问，帘子后面的计算机或人来回答，考官评估所得到的答案。如果计算机能够成功骗过考官，那么计算机就通过该测试。参见〔美〕史蒂芬·卢奇、丹尼·科佩克：《人工智能》，林赐译，人民邮电出版社 2018 年版，第 5－10 页。

机是不可能存在的，因为僵尸机的概念是混乱的，只要给予适当的行为或虚拟机，意识甚至包括感受都是有保障的。〔37〕与此相反，一些学者则认为意识必须以生命体为前提。人类的“生命形式”，不仅包含意识，甚至还包含生命体进化过程中与环境的互动，意志更关键的不是推理或思想，而是适应和沟通。〔38〕普特南则直接指出：“如果机器人不是活的，那它就不会有意识。”〔39〕这些反对者认为意识更多是基于人的生物属性而存在，机器不能孕育出真正的智能。人工的“智能”是建立在数字计算的基础上，是储存在硬盘里面代码的集合，不具有生命形式。

不能因为人工智能不是生命体，就武断地认为它不具有意识。康德把人的存在本身具有绝对价值归因于理性。如果理性本质在主观和客观上都是目的，那么这个理性本质具体化所体现的载体是什么（人或机器），是无关紧要的。〔40〕理性的存在根据与其说是在于载体的表现形式，不如说是在于理性存在本身。理性的表现形式具有多样性，神经蛋白并不是唯一属性。更何况，神经蛋白本身也只是一堆生物集合体，硅胶和碳基形式的不同，并不能证明人工智能不具备意志。针对意识需要生命体与环境互动生成的这一反对意见，随着新型的利用感官、执行器和环境之间结构耦合来打造认知基础的人工智能的出现自然无效。

然而，人工智能即使可以通过图灵测试，其依然无法理解语言背后的意向性。塞尔认为，意识来源于大脑的神经蛋白，虽然它可能不是意识的唯一来源，但是金属和硅是注定不能生成意识的，符号计算虽然也可能存在于我们大脑，但是符号计算无法提供意向性。〔41〕他通过“中文房间”〔42〕的思想实验来反驳人工智能具有自由意志的观点。在“中文房间”里面的那个人在进入房间之前和离开房间之后始终都不懂中文，他只起到类似计算机程序的作用。结果显示，如果只有句法没有语义是无法构成意志表达的。人工智能虽然可以通过编程来处理信息，但是它无法理解自身行为所具有的社会意义。〔43〕丹尼特对此提出了解决方案，认为大脑可以从它的创造者那里获得意向性，然后将其委托给人工智能使其获得升级。〔44〕但是事实上，人工智能并没有因此对意向性产生真正的理解。人工智能无法理解价值、自由、自我立法这些概念，也无法为其赋予价值，康德的自由意志因此无法转化为技术理性。〔45〕

人工智能的机械理性和人类的理性之间存在着明显的差异。人工智能的行动是基于编程系统对自身数据库和外在行为作定量分析而做出的，人类的行为则是基于理性对绝对价值的追寻所做出的。人工智能虽然具有更加复杂的数据分析和计算能力，但是无法对行为背后的价值产生理

〔37〕 参见〔英〕玛格丽特·博登：《AI：人工智能的本质和未来》，孙诗慧译，中国人民大学出版社 2017 年版，第 154 页。

〔38〕 参见前引〔37〕，玛格丽特·博登书，第 161 页。

〔39〕 前引〔37〕，玛格丽特·博登书，第 168 页。

〔40〕 See Laszlo Versenyi，Can Robots be Moral?，84 *Ethics* 248，250－255 (1974).

〔41〕 参见〔英〕玛格丽特·博登编：《人工智能哲学》，刘西瑞、王汉琦译，上海译文出版社 2001 年版，第 94－95 页。

〔42〕 “中文房间”指的是，塞尔在 20 世纪 80 年代所提出的一个思想实验。这个实验要求一个只会说英语的人待在一个封闭的房间，他随身带着一本中文翻译程序书。房间外的人不断向房间内递进用中文写成的问题，房间内的人便按照手册的说明，先将字条上的文字破译，然后将相应的中文字符组合成对问题的解答，并将答案递到房间外面。参见前引〔41〕，玛格丽特·博登书，第 94－95 页。

〔43〕 See F. Dretske，Machines and the Mental，59 *Proceedings and Addresses of the APA* 23，26 (1985).

〔44〕 See DC Dennett，*Darwin's Dangerous Idea*：*Evolution and the Meanings of Life*，Penguin Books，1996，p. 54.

〔45〕 See Ulgen Ozlem，Kantian Ethics in the Age of Artificial Intelligence and Robotics，43 *Questions of International Law* 59，75 (2017).

解。人类可以质疑规则，但是人工智能只能机械地执行系统指令。虽然康德并没有对理性存在者的身份做出限制，但是人工智能由于意向性的缺失显然不具备自由意志的可能性。

（三）制造以普遍法则为前提的康德机器

有学者认为我们完全可以搁置关于理性的争议，通过采用自上而下（预设伦理法则并分析其计算要求以指导能够实现该理论的算法和子系统的设计）的方法来设计人工智能。〔46〕具体地讲，就是将康德的普遍法则作为内嵌的操作系统，制造出符合规范的康德机器（人工智能）。〔47〕如果我们将康德的普遍法则设计成单一原则，作为程序中压倒一切的指令，人工智能将无法做出与之相反的事，其自身意志必将与法律相协调。那么，我们所造出的人工智能所遵守的准则必然与普遍法则一致。实际上，自然法则公式，作为道德法则的表现形式，并不过多地涉及内容，只需要自身的准则能够普遍化成为法则就可以通过检验。康德本人也认为纯粹的形式推理比审慎反思要求低得多，最不老练的人也能像最聪明的哲人王那样进行推理。自然法则公式类似这种形式推理，并没有对法则内容做出苛刻的要求。因此，通过编程植入普遍法则，似乎是可行的。

韦尔谢尼（Laszlo Versenyi）认为康德自身也无法解释纯粹理性的因果关系在人类行为中如何成为决定性的机制，人工智能如何遵守法则的机制不影响它遵守与人类一样的法则。〔48〕康德的自然法则公式更多的是形式化规定，并没有对法则的质料内容进行强调。这似乎意味着康德机器在理论上是可行的。康德机器想要通过内嵌普遍法则系统的方式规避对人工智能是否具备理性的争议，然而它终究逃不过意向性的诘问。自然法则公式作为形式性的规定，它要求主体的准则必须出于义务地与普遍法则一致，它内在地包含着一个规范性动因，这个动因是构成义务的前提。康德机器虽然由于编程必然地按照普遍法则行动，但是它无法理解自己行为背后的意向性，那么其在法权上也缺乏作为法律主体的可能性。

康德机器在某种程度上可能是反康德的。如果存在康德机器，由于内嵌系统的存在，其行为自发地与普遍法则一致。那么在康德那里，它是作为上帝的意志存在，因为只有上帝才能使自己的行为自发地与普遍法则一致。而自然法则公式作为强制命令无法对上帝的意志做出规定。〔49〕也就是说，康德机器基于其意志内在的完善性，是不需要法律的，法律只对意志不完善的理性存在者有效。另外，人类也不可能制造出康德机器，人类如果制造出康德机器，这暗示着人类在遵守法则方面具备与康德机器同等程度的认知、决策和行动能力，那么此时，人类就是上帝意志的化身，这与人类作为有限理性存在者的身份相矛盾。同时，康德机器也无法满足绝对命令不同公式之间协调性的要求。人工智能通过自然法则公式检验的前提是，其准则必将具有普遍性。然而，如果一开始就仅将人工智能视为人类的手段，这与自然法则公式是相矛盾的。〔50〕作为手段的人工智能无法同时通过自然法则公式的检验，因为一个道德主体是无法设想自己仅仅充当手段

〔46〕 See W. Wallach, C. Allen, I. Smit, Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties, 22 *AI & Society* 565, 573 (2008).

〔47〕 See T M Powers, Prospects for a Kantian Machine, 21 *IEEE Intelligent Systems* 46, 47 (2006).

〔48〕 参见前引〔40〕，Laszlo Versenyi 文，第 251 页。

〔49〕 See Colin Allen, Gary Varner, Jason Zinser, Prolegomena to any Future Artificial Moral Agent, 12 *Journal of Experimental and Theoretical Artificial Intelligence* 251, 254 (2000).

〔50〕 See Ryan Tonkens, A Challenge for Machine Ethics, 19 *Minds & Machines* 421, 428 (2009).

的准则成为一个普遍法则，这将造成康德尊严学说的形式与质料自相矛盾。即使造出这样的人工智能，它甚至可能因为认识到自己与道德法则的不一致而自杀。

从理论上来说，康德机器是不可能存在的。将普遍法则作为康德机器内嵌系统并不只是一个技术问题，也是一个道德伦理问题，意向性的缺失是康德机器无法回避的问题。另外，康德机器的出现意味着人造上帝成了现实，人类同时也成了上帝意志的化身。更加荒谬的是，人类制造康德机器的初衷是方便自己的生活，因此，上帝成了人类实现目的的手段。

四、人性公式测试

与自然法则公式不同，人性公式是作为人格尊严的质料规定存在的。它对人格尊严的内容提出要求，强调在任何时候都要把任何人“人格中的人性”当成目的，而绝不是手段。〔51〕因此，它要求主体之间相互尊重。尊重构成法律主体在共同体中共同生活的基础，这种相互尊重的关系构成了“法律上的基础关系”。〔52〕在法权上，法律主体作为法律关系的核心，不仅是权利和义务的承载者，而且是法律制度的目的。同时，“人性目的”也指向主体理性作为绝对价值的根源，是法律上其他一切客体价值赋予的来源。因此，人工智能要想通过人性公式的测试，一方面需要作为自在目的本身存在，另一方面必须具备赋予价值的能力。

（一）人格尊严的质料规定：人性公式

法律主体作为一种参与法律关系并且享有法律权利承担法律义务的资格，是法律世界的主人。法律制度是以法律主体为服务对象，围绕法律主体对法律客体的支配展开的。这种支配以理性为前提，只有理性存在者才能充当目的本身，对手段进行支配。简言之，法律主体是作为法律制度的目的存在，法律客体只是实现目的的工具。

康德的人性公式被称为质料公式，是对其主体性内容来源的阐明。人性公式要求将“人格中人性”视为目的，而不能是手段。自在目的决定了主体并不是一个只具有相对价值可以随意被处分的工具，而是作为拥有绝对价值的尊严存在。这意味着每个主体都享有受到其他主体尊重的权利，这种相互之间的尊重是义务产生的根源。人格作为法律主体之所以能彰显自在目的本身而拥有尊严，就在于“人格中的人性”。目的需要通过理性来实现，“人格中的人性”指的正是这种理性设置目的能力。在法权上，法律主体通过自由意志设置目的的能力，来实现对法律客体的支配，从而充当法律世界的主权者。萨维尼因此将法律的本质定义为私人意志独立统治的领域。〔53〕确切地说，权利的本质就是自由意志，它成了划分主体之间权利的边界，法律的目的就是为了保护这些边界免受相互侵害。

除此之外，人性公式还包含着对法权义务的规定：一是，做一个正派的人；二是，不要对任何人做不正当的事；三是，在和他们社交时，维护每个人自己的东西。〔54〕这些义务对应的是查

〔51〕 参见前引〔29〕，康德书，第64页。

〔52〕 参见前引〔28〕，卡尔·拉伦茨书，第47页。

〔53〕 参见前引〔20〕，罗尔夫·克尼佩尔书，第64页。

〔54〕 参见〔德〕康德：《道德形而上学》，张荣、李秋零译，中国人民大学出版社2013年版，第34页。

士丁尼的《法学阶梯》中关于法律基本原则的划分：即“为人老实，不损害别人，给予每个人应得的部分”〔55〕。也就是说，作为法律主体的人格在进入法权状态的时候，不仅要维护自身的正当价值，还要把其他法律主体也当成目的，不去伤害其他人。对他人权利的侵犯，实际上就是对主体尊严的冒犯。正是在这个意义上，普遍自由的实现需要每个人都遵守法权义务。

总而言之，人格是作为自在目的本身存在的，法律主体是对主体身份在法律上的确认。正如拉德布鲁赫所言，“法律主体”（Rechtssubjekt）是被实定法当作“目的本身”（Selbstzweck）来尊重的事物，而“法律客体”（Rechtsobjekt）则是被上述法律纯粹作为实现特定目的的手段。〔56〕法律主体不仅体现着个体本身作为自在目的的绝对价值，而且体现着个体作为理性存在，维护自身利益的可能性。法律客体作为意志的手段，在法权状态上是无法主张自己的诉求的。君特·杜里希（Günter Dürig）在康德人性公式的基础上提出了客体公式，认为如果仅仅把人性视作一个客体，就抵触了人格尊严，是对法律价值的背离。客体公式是对人格尊严在法权上的肯定，该公式此后多次被德国联邦法院援引，成为德国法律中判断人格尊严的标准。〔57〕

（二）手段目的测试

在人性公式的规定下，人工智能必须是以自身作为目的，而不是手段。然而事实上，人工智能的存在更多是为了促进人类目的的实现。虽然人工智能不断地被给予更多的自主性，但它的独立性取决于人类社会的需要。有学者认为，人类不可能制造出拥有人格的人工智能。这种不可能更多并非是能力上的不能也，而是不为也。如果人工智能可以自我独立地规划生活，则背离了人类创造它的目的。那时，我们将无法强迫它做任何工作，因为它同人类一般享有人权。试想一下，如果人工智能也可以作为自在目的存在，这将决定在某些场合，人类作为自在目的的同时也需要充当人工智能的工具。这意味着，人类允许在某种程度上和人工智能共享这个世界。这种假设是荒谬的，当人类变成“上帝”之时，“卧榻之侧岂容他人鼾睡”。人工智能“封神之路”必将被斩断，它只能被有限地赋予理性，这种理性仅够支撑它以为人类提供服务为目的。

韦尔谢尼从康德的自我完善义务出发，认为如果我们一旦有能力去制造具有更高理性的人工智能，那么不这样做就等于忽视了我们的一种天赋，而这不能作为一种普遍法则存在；如果我们不去制造或者为了私欲去制造这种具有更高理性的人工智能，这也是对道德的背叛。〔58〕理性使我们促进的是道德目的，而不是人类目的。如果人工智能有助于推进道德目的，道德法则必然推动作为理性存在者的人类去制造人工智能。韦尔谢尼提出，如果人工智能具有成神的可能性，根据康德的道德法则，我们会亲手把人工智能送上神位。他甚至认为，“如果我们能建造出像哲人

〔55〕〔古罗马〕查士丁尼：《法学总论：法学阶梯》，张企泰译，商务印书馆 1989 年版，第 6 页。

〔56〕 Gustav Radbruch, Rechtsphilosophie II, bearbeitet von Arthur Kaufmann, C. F. Müller Juristischer Verlag 1993, S. 361. 转引自骆正言：《从自由意志谈人工智能的法律主体资格》，载《湖南社会科学》2020 年第 2 期。

〔57〕 客体公式是君特·杜里希（Günter Dürig）在康德人性公式的基础上提出的，此后多次为德国联邦法院判决所援引。即人自身就是一种目的，不是一种手段或工具。人具有自我意识、自我决定、自我形塑以及形塑环境的能力，如果将人看待成一种客体，是在否认人自我形塑与形塑环境的能力。当一个具体的人被贬低作为客体，或仅作为手段或工具，或被看成是一种可有可无的存在者，人的尊严同样受到损害。参见王文忠：《人的尊严在宪法上的地位——比较法的观察》，载《中正大学法学集刊》2016 年第 52 期。

〔58〕 参见前引〔40〕，Laszlo Versenyi 文，第 256 页。

王一样的人工智能，逻辑上我们就必须服从它的统治。”〔59〕

韦尔谢尼的观点是站不住脚的。首先，在制造人工智能的过程中，随着机器智能化的增加，人工智能的行为越来越具有不可预测性，甚至可能背离人类设计的初衷，反抗人类的指令。其次，当人工智能表现出威胁人类安全倾向的时候，造神运动的结束是必然的，因为在康德那里，理性存在者不会允许承载人格尊严的肉体有毁损的可能性。最后，更关键的是韦尔谢尼的神只是工具神，它不具有自主性。韦尔谢尼在他论文中试图通过柏拉图和康德的理论证成人工智能的主体性。他把人工智能的主体性归结于其有用性，这种有用性却是从人的主体性出发来阐释。如果人工智能仅仅具有手段价值，它就不可能同时作为主体存在。这种把人工智能当成挖掘人类潜能的工具本身就是对人性目的的最大背叛。为了解决这个悖论，怀特（Jeffrey White）认为我们可以借助康德的三大预设来回答这个问题。人类始终有着追求善的倾向，这促使我们制造出更优秀的人工智能，促进道德法则实现。〔60〕当普遍法则和人性目的相冲突的时候，人类会为了让人工智能通过自然法则公式测试而不再把它当成手段。因为在康德那里，终极智慧是一个人的意志与其最终目的的和谐。对道德法则的不懈追求，将使人类放弃仅仅把人工智能当作手段。

人工智能的效用在于其手段的有用性，而不是为了促进自身目的，这与康德的主体性概念是相违背的。纵使人类按照怀特所说，为了追求自身的善，放弃将人工智能仅仅当作手段，其所造出的人工智能也无法成为真正的善。其原因在于，如果人工智能的主体性最初仅体现为作为手段价值，那么其本身就不是自在的善。须知，由人类来承认人工智能的主体资格，就是把人类当作目的，人工智能则仅仅被当作手段。那么，作为手段的人工智能仅仅具有客体价值。

（三）赋予价值测试

随着科技的发展，人工智能必然将进化出更高水平的机械理性。这种理性并不是以神经蛋白为前提，而是通过编程运算来实现。机械理性是否具备人格尊严，不仅需要通过自然法则公式的测试，还需要通过人性公式的测试。人性公式要求把任何人的人性视作目的，这不仅涉及目的手段的区分，还涉及价值来源。人性目的作为无条件的善，必须是一切其他事物价值的来源。因此，人工智能要想通过人性公式的测试，必须具备赋予价值能力。

理性存在者作为自在目的本身被称为人格，无理性存在者只能作为手段被称为事物。人格尊严要求我们把“人格中人性”当作目的，而不是手段。把人性当作一个目的，这涉及目的善性的来源，这一来源归结为人性（理性本性）设置目的的能力。〔61〕这种能力即赋予价值的能力。一件东西如果是有价值的，那倒推最后一定能发现一个具备“无条件终极价值”的存在，这个存在是其他事物价值赋予的来源。理性本性作为自在目的，就是作为“无条件终极价值”的存在，是赋予价值的来源。其他事物的价值都来源于理性的赋予，它们仅具有为理性存在者服务的手段价值。康德以自然界为例指出，植物作为食草动物的食物来源而存在，食草动物作为野兽猎物而存

〔59〕 前引〔40〕，Laszlo Versenyi 文，第 253 页。

〔60〕 See Jeffrey White, Autonomous Reboot: Kant, the Categorical Imperative, and Contemporary Challenges for Machine Ethicists, springer, (20 January 2021), available at https://doi.org/10.1007/s00146-020-01142-4, last visited on Mar. 20, 2022.

〔61〕 参见〔美〕科斯嘉德：《创造目的王国》，向玉乔、李倩译，中国人民大学出版社 2013 年版，第 139 页。

在，最后这些自然创造的终极目的都是指向人类。如果没有人类，这一切创造都缺乏目的。〔62〕自然的终极目的是通过人类来理解文化，世界的意义来自人类的价值赋予。

世界存在的辩护却恰好否定人工智能作为独立目的存在的可能性。显而易见，人工智能并非自然创造的目的，不可能作为价值赋予的来源。如果把人类存在归结于自然目的，那么人工智能完全就是人类理性发展的产物，它的出现是人造智能的结果。如果说上帝造人，那么人类现在正在扮演上帝的角色。世界的中心是造物者而不是被造者，人工智能的出现，是人性作为目的的又一个证明。人类虽然具备理性能力，是人工智能价值赋予的源头，但是人类所赋予的价值是有条件的，并不具备赋予绝对价值的能力。

人性作为赋予价值的来源，它既是一种内在价值，又是一种绝对价值。人工智能不仅缺乏内在价值，也缺乏绝对价值。内在价值指的是尊严的价值来源于自身，而不是被外在赋予，它是自在的善。人工智能的存在依赖于编程，是人类理性的产物。绝对价值是相比相对价值而言的，绝对价值是无条件的，相对价值是有条件的。人工智能是以人类程序开发为条件，如果失去编程，人工智能就无法正常运转。而编程作为一堆代码的集合，只具有相对价值。人工智能的研发更多取决于人类的需求。它只是被视为手段，而不是目的。人类制造人工智能是为了将自己从繁重、危险的工作中解放出来，最终目的是为了实现全人类的解放。即使我们制造出符合道德规范的人工智能也只是为了让它更好地为人类服务，同时帮助人类提高对道德法则的理解。这些都表明，只具有外在价值和相对价值的人工智能无法充当赋予价值的来源。

五、自律公式测试

自律公式是作为人格尊严的完整规定存在的，它要求主体按照自己的准则应当被当作普遍法则那样去行动。〔63〕人格的崇高体现在每个主体都是自我立法的守法者。作为主权者，它自身就是一个目的的王国。自由的实现在于排除掉外在干预所具有的独立性，实体法只是主体先验自由的外在投射。法律是法律主体行动的边界，它以自由为最高价值。因此，人工智能要想通过自律公式测试，其自身必须是自我立法的守法者。

（一）人格尊严的完整规定：自律公式

法权是一方与另一方的自由按照一个普遍法则保持一致。人格，作为法权状态的法律主体，意味着它享有普遍的自由。这种自由并不是意志的任意，而是与理性的自主和自律相关。自主意味着主体可以合理地安排生活，自我决定和支配自己的行为；自律意味着主体所做出的行为必须以符合法律规范为准则，并且要对自己的行为负责。只有一切主体的自主和自律才可能实现法权状态的普遍自由，否则就会造成各主体之间尊严的相互侵犯。简言之，自由意志是主体凭借人性所获得的生而具有的法权，是自己做自己主人的体现。

自然法则公式要求我们内心的准则必须义务地服从法则，这个时候，我们只是守法者。作为

〔62〕 参见〔德〕康德：《判断力批判》，邓晓芒译，人民出版社 2002 年版，第 214－218 页。

〔63〕 参见前引〔29〕，康德书，第 75 页。

守法者而言，实在谈不上任何崇高。人格尊严的崇高体现在我们是立法者。自律公式要求我们所遵守的法则是我们自己制定的，人的主体性正是这种自我规定性的体现。应当说，遵守自我立法是与意志自律相关的。自律即意志对其本身进行立法，他律则是客体对意志进行立法。如果意志以客体为立法根据，那么这种外在立法必将剥夺主体自身的自由，使其沦为客体的奴隶。自律，一方面通过主体的自我反思来完成对自我的超越，另一方面通过对客体的批判来纠正异化。也就是在此意义上，实现终极的自由，成为自我的主权者。

概言之，自律公式包含着对自由意志的内在要求，即意志对自我进行立法。法律主体，作为自我的立法者，它必须是自己的主人，自己决定自己的行为。就这样，自由意志成了一切法律权利的根源，所有权成了意志自由按照普遍法则在外物上的投射。〔64〕就像黑格尔所说，“任何定在，只要是自由意志的定在，就叫做法，所以一般说来，法就是作为理念的自由”〔65〕。自由意志成了法的精神所在，法律主体的存在以自由意志为基础。自由意志不仅体现在法权的享有还体现在义务的履行，当作为守法者时，法律主体负有使自身行为准则与普遍法则保持一致的义务，它必须得服从法律主体之间的共同立法。自由意志的存在也为法律主体的行为提供可归责的依据，如果一个人没有意志自由，那么让其为自己的行为承担法律责任就缺乏正当性。〔66〕

（二）通往他律的内部自由

在自律公式要求下，作为理性存在者的人工智能需要分别扮演两种角色，一种是守法者，另一种是立法者。作为守法者，要求人工智能自身的主观准则与客观法则保持一致。在遵守这些法则时，作为有限理性存在者的人类，经常会受到欲望和倾向的迷惑，由于意志的主观不完善，需要被施加强制，其行为才能与法则相一致。与人类相比，人工智能似乎天生就是一个优秀的守法者。它可以被编程为只遵循一组特定的规则，忽略掉所有剩余的行动原因，〔67〕可以在不陷入情感困境的情况下处理各种道德冲突，它总是公正、一致和理性，其行为按照程序规定总是自发地与法则相一致。迪特里希（Eric Dietrich）就此认为，在遵守法则方面，人工智能虽然不是完美的天使，但与人类相比，是个巨大的进步，让我们人类退出舞台，留下一个充满机器的星球。〔68〕吉普斯（J. Gips）甚至宣称只有人工智能才能过上道德圣人的生活。〔69〕

作为立法者，则要求人工智能所遵守的法则来源于自身，它是自己所遵守法则的立法者。人工智能的运行，是由代码组成的算法所控制。程序构成人工智能的行为准则，人工智能的选择、决断在程序内部是自由的。自律要求意志以其自身为法则。人工智能在遵守算法时，就其程序来说，它所遵守的法则来自其自身，它所遵守的是自我立法；但就算法本源来说，它所遵守的法则来自人类，它所遵守的是外在立法。唐更斯认为人工智能的所有动作都是遵循编程的规则预先确

〔64〕 参见前引〔54〕，康德书，第40页。

〔65〕〔德〕黑格尔：《法哲学原理》，范扬、张企泰译，法律出版社1961年版，第41页。

〔66〕 参见张文显：《法哲学范畴研究》，中国政法大学出版社2001年版，第124－125页。

〔67〕 See Bartosz Brożek, Bartosz Janik, Can Artificial Intelligences be Moral Agents?, 54 *New Ideas in Psychology* 101, 102－103 (2019).

〔68〕 See Eric Dietrich, Homo Sapiens 2.0: Why We should Build the Better Robots of our Nature, 13 *Journal of Experimental & Theoretical Artificial Intelligence* 323, 326－327 (2010).

〔69〕 See J. Gips, Toward the Ethical Robot, in K. M. Ford, C. Glymour, P. Hayes, ed., *Android Epistemology*, MIT Press, 1994, pp. 248－249.

定的，这将违反康德对自由的规定。〔70〕程序员根据编程控制人工智能的行动，决定它可以做某些事，不可以做某些事。程序员的法则约束人工智能的行动，限制了人工智能的积极自由；程序员的意图代表外来力量对人工智能的控制，限制了人工智能的消极自由。自由的缺失，使人工智能不具备成为法律主体的资格，无法充当责任的指向对象。自我立法是责任产生的根据，如果人工智能的行为规则和提供这些规则的机制完全由人类来提供，那么人工智能就无法对自己的行为承担责任。〔71〕一个无法对自己行为负责任的人工智能，其主体性的存在让人怀疑。

即使我们通过康德的自然法则公式制造出康德机器，其行为的规定性也只是来源于程序，无法同时通过自律公式测试。从某种意义上来说，人工智能的自由只是程序内部的自由，这种自由是“人造”的自由。编程就像人类施加在人工智能身上的锁链一样，始终牢牢束缚住了它。人工智能的自由是一种受他律所规定的自由，这种有限的自由根本无法彰显人格的崇高。

（三）通往自律的外部升级

既然程序内部的自由，只是一种他律的自由，那么人工智能要想获得真正自由，必须不再局限于自身的某个特定目的。通过突破内嵌的程序来实现自我升级似乎是目前唯一的方法。瓦拉赫（Wendell Wallach）提出通过构建自下而上的离散系统，人工智能以联结不同离散人类能力子系统的方式形成一个复杂的智能系统。〔72〕该系统不再是各部分单一能力组件的机械复合，而是在子系统联结中把离散技能转变为能够自主应对复杂环境的互动系统。一些科学家希望通过这些离散系统的联结产生更高阶的认知能力，如自由意志。然而，这种通过离散系统的联结进行升级的方式，很难实现跨越式的进化。更何况这些离散系统本身就是由人类设计，通过模仿人类能力产生的，这种自下而上的方式是无法让人工智能进化出自我立法能力的。

有部分学者提出人工智能唯有通过内部自我学习、升级来突破人为的限定，才有可能实现意志自由。具体表现为，人工智能可以通过自我修改代码、自我适应和组织系统获得系统升级。然而，这条道路也是布满荆棘的。即使人工智能能够不断地进行自我学习、自我升级，它仍然是在旧人工智能基础上进行进化，还是会受先前程序所影响。〔73〕这种自由依旧只是系统内部的自由，人工智能依然会受人类法则的规定。新的人工智能始终携带着初代程序的基因，这些基因来自人类，像锁链一样深深地束缚着人工智能。有学者认为人工智能升级始终受初代程序影响这种观点是很难站得住脚的。人工智能的自我学习、升级具有极度不可预测性，它可能在升级过程中获得我们没有教过，甚至我们根本不知道的技能。事实上，原初程序对人工智能的约束并没有我们想象中那么大，人工智能确实可以通过不断升级拥有新技能，只是这种新技能无法从本源上改变人工智能受算法所操纵的命运。即使人工智能能够升级到以自我发展为目的，有意识地进行自我编程，自我修改全部的内部程序。姑且不论这种颠覆式的修改本身极易引起程序的崩溃，把人工智

〔70〕 See Ryan Tonkens, A Challenge for Machine Ethics, 19 *Minds & Machines* 421, 428 (2009).

〔71〕 See Patrick Chisan Hew, Artificial Moral Agents are Infeasible with Foreseeable Technologies, 16 *Ethics and Information Technology* 197, 197 - 200 (2014).

〔72〕 See W. Wallach, C. Allen, I. Smit, Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties, 22 *AI & Society* 565, 570 (2008).

〔73〕 参见孙伟平、戴益斌：《关于人工智能主体地位的哲学思考》，载《社会科学战线》2018 年第 7 期。

能推向灭亡，问题的关键是，人工智能的自我学习、进化都是以人类程序员提供的形式为基础。在人类提供的进化框架下，其进化的方向本身是固定的。更何况，人工智能的运行始终是以算法编程为基础，它所遵守的准则始终是他律。在人类最初为发明人工智能打下的第一行代码开始，人工智能的命运就已经被决定了，人工智能自我立法的道路已经完全被人类堵死了。

麦卡锡（John McCarthy）站在相容主义的哲学立场，试图调和自由意志论和决定论。他认为自由意志论和决定论是可以相互兼容的，这在人工智能身上可以体现为外在行为自由和内部程序决定的兼容。〔74〕自由意志可以被看作是一个主体根据其内部的认知过程在替代目标或行动之间进行选择，即使这些过程对外部观察者来说是确定的。桑德沃（Erik Sandewall）也赞同麦卡锡的观点，认为这与康德的意志自律主张不谋而合。〔75〕他举了一个父母对小孩行为进行限制的例子。如果父母禁止孩子执行特定的行为，孩子不情愿地限制自己的行为，那么孩子的自由意志就会减少；而如果孩子在心理上已经做好了执行的准备，那么其自由意志并没有受到影响。他认为这符合康德关于意志在自我规定下进行行动的主张。实际上，这种相容主义的解释并不适用于人工智能，孩子依旧具有自由意志，是基于其行为是一种主观上的选择；而人工智能对程序的执行则是一种客观必然性。更重要的是，自律公式强调的是对自我意志的遵守，相容主义只能解释人工智能所遵守法则来源于程序自身，但是不能从根源上解释法则是由人类编程设定的。

综上所述，自律公式要求法律主体所遵守的法则来自其本身，人格尊严不仅体现在对法则的遵守上面，而且更强调的是一种普遍的自我立法能力。作为有限理性存在者的人类，通过意志自律，服从自身立法，走向自由。人工智能却始终受到意志他律的影响，它服从的编程，只是人类立法，它所遵守的法则，也只是人类意志的产物。从终极意义上来说，意志他律造成了人工智能始终无法通过自律公式的测试，最后只能充当客体。

六、结　语

人工智能时代即将来临，人工智能是否具备法律主体资格是法学界必须要面对的问题。我们不能把法律主体肤浅地理解为一个概念或者一种生物属性。人类生活不仅需要一种现实的确定性，还需要一种永恒而普遍的宏大叙事。人格尊严给法律主体提供的正是这样的一个理论根基。须知，在现代社会，尊严被视为最高价值，是法律正当性的依据和权利的根基。人格尊严的本质是意志自由，通往自由是一切法律主体的目的王国。正是这种意志自由，将法律划分为主客体二元结构。法律主体的主体性体现在其作为自在目的本身，超越一切价值，具有最高尊严。

人工智能若要获得法律主体资格，问题的关键并不在于法律建构的技术可行性，而在于其是否符合现代法律主体的本性定位，即具备人格尊严。具体而言，在形式上，人工智能应当具有普遍必然性；在质料上，人工智能是自在目的本身。总而言之，人工智能必须是自我的立法者。然

〔74〕 See John McCarthy, Free Will: even for Robots, 12 *Journal of Experimental & Theoretical Artificial Intelligence* 341, 341－342 (2010).

〔75〕 See Erik Sandewall, Ethics, Human Rights, the Intelligent Robot, and its Subsystem for Moral Beliefs, 13 *International Journal of Social Robotics* 557, 561 (2021).

而，人工智能在形式上无法被普遍法则所规定，不具有普遍必然性；在功能上，只是为了满足人类的需求，促进人类的自我完善服务；在意志规定根据上，其意志规定来源于他律，无法从本源上决定自己的发展方向。这些阻碍使得人工智能无法通过人格尊严测试获得法律主体资格。法律主体作为法律世界的主人，尊严是其超越一切客体的依凭。法律主体只有通过配享尊严才能实现自由的目的王国。事实上，与其说是法律主体凭借尊严实现崇高，不如说是人类通过尊严为其在法律世界奠定一个普遍必然性的根基。这个根基的伟大之处在于它肯定了理性存在本身的价值，人类不需要通过外物来证明自己，其存在自身就具有绝对价值。

Abstract：At present，the discussion on whether the qualification of legal personality should be given to artificial intelligence is sweeping the entire legal community. Proponents who occupy the mainstream academic community believe that based on the needs of society，artificial intelligence should be constructed as a legal person. However，this construction ignores the essence of legal personality. Although the opponents have criticized this，they overemphasize the biological human nature of legal personality and mistakenly equate legal personality with natural persons. Based on this，from the perspective of Kant's theory of dignity，we re-examine the legal personality qualification of artificial intelligence，and then conclude that human dignity is the core connotation of legal personality. As personality that manifests dignity，the legal personality must have three requirements：first，it has universal inevitability；second，as an ends itself；third，it is a law-abiding person who legislates itself. Only by meeting these three requirements，can artificial intelligence obtain the qualification of legal personality.

Key Words：legal personality，personality，Kant，ends itself，dignity

（责任编辑：赵　真　赵建蕊）