边沁、罗尔斯与分配正义的算法化研究

徐英瑾*

〔摘要〕 目前关于人工智能运用的主流制约性规范，往往受到康德伦理学的影响，但是学界对于相关的规范伦理学立场的可算法化问题却一直没有系统的检讨。实际上，无论是边沁的功利主义理论，还是罗尔斯的分配正义理论，甚至德性论—社群主义关于社会资源分配的思想资源，其实都无法成为某种可以被算法化的思想指导，以便引导编程人员最终完成社会资源分配机制的全面自动化。这一消极的结论无疑会给基于任何一种规范伦理学立场的人工智能伦理学方案蒙上阴影，但也由此为人类主体在未来社会决策中所扮演的角色预留了足够的空间。

〔关键词〕 功利主义 边沁 罗尔斯 人工智能 算法 电车难题

〔中图分类号〕B82-057 〔文献标识码〕A 〔文章编号〕1007-1539(2022)05-0159-11

DOI:10.13904/j.cnki.1007-1539.2022.05.008

一、导论

人工智能产品在人类生活中日益广泛的应用，不可避免地引发了学界关于其伦理学后效的广泛思考。由此类思考所激发的分支学科，即“人工智能伦理学”(ethics of Artificial Intelligence)。早期的“人工智能伦理学”是以“机器人伦理学”(robot ethics)的面目出现的，其典型的思维结晶便是科幻作家阿西莫夫的“机器人三法则”。但随着人工智能的发展，学界日益注意到人工智能未必一定要采取机器人的物理外观，因此，“机器人伦理学”这个名目的重要性日益被让位给“机器伦理学”(machine ethics)这个覆盖力更广的新名目。需要指出的是，与技术伦理学的其他分支一样，人工智能伦理学往往是某种更为一般的规范伦理学立场的具体化。譬如，阿西莫夫的“机器人三法则”其实就是康德的人本主义伦理学的某种改写版；无独有偶，欧盟在新世纪制定

* 作者简介：徐英瑾，复旦大学哲学学院教授、博士生导师(上海 200433)。

的种种数据规范法则亦具有鲜明的康德伦理学的人本主义色彩(譬如,2018 年 12 月 18 日欧盟委员会公布了《可信赖的人工智能道德准则草案》[①],根据其精神,人工智能产品的研发必须以人为中心,尊重人类的尊严、平等和自由等基本权利)。然而,由此进入机器伦理学视野的规范伦理学立场,往往并没有被相关的技术规范立法者所关注,而成为某种缄默的预设。这在一定程度上亦影响了此类立法研究的思想深度。

本文的立论将摆脱现有的机器伦理立法工作的思想局限,而试图思考一个更具前瞻性的话题:在未来某个时刻,如若我们需要用 AI 来协助人类对于社会资源(如医疗资源、教育资源等)的分配的话,那么,我们需要何种伦理学规范作为相关的编程作业的思想指导呢? 依据笔者的猜测,在面对上述问题时,很多人或许都会诉诸罗尔斯的分配正义理论,因为罗尔斯的理论既在思想脉络上继承了作为主流机器伦理学研究之思想圭臬的康德伦理学,同时又具有康德伦理学所缺乏的"算法细节"。此外,罗尔斯的分配正义理论已经通过其社会影响而渗入了一些西方国家的民意代表结构的立法程序。因此,至少在西方国家的范围内,该思想已经具有了一定的政治权威性。从这个角度来看,如果某些国家的立法者试图对社会资源分配的流程加以程序化或者自动化的话,由此所催生的成果就很可能是"罗尔斯式的"。

但本文则试图指出,罗尔斯式的社会资源分配方案是无法被真正算法化的,而富有讽刺意味的是,罗尔斯的方案无法被算法化的理由,恰恰便是作为其主要理论对手的功利主义的社会资源分配方案也无法被算法化的理由。另外,笔者还将论证:作为功利主义与义务论(康德与罗尔斯均属于该阵营)之外最重要的规范伦理学立场,德性论—社群主义的进路所能够提供的"可被算法化"的空间也是有限的。所以,笔者的结论是:几乎没有一种主流的规范伦理学立场可以被彻底地算法化。这一消极的结论无疑会给基于任何一种规范伦理学立场的机器伦理学方案蒙上阴影,但也由此为人类主体在未来社会决策中所扮演的角色预留了足够的空间。

我们下面的讨论将以电车难题为引子,并由此切入边沁式的社会资源分配方案。

二、对于功利主义的社会资源分配方案的再考察

众所周知,电车难题是战后英美伦理学用以考察义务论与功利主义进路之彼此短长的一个经典案例[②]。这一案例亦是社会资源分配的一种极端情形:在此,被分配的资源不是别的,而是利益相关者的存活机会。在本节中笔者将在自动化驾驶的语境中对这一案例再进行重述。

现在假设有一辆自动驾驶的有轨电车正在向前行驶。此刻,电车的红外传感器突然发现电

① European Commission(2018),High-Level Expert Group on Artificial Intelligence:Draft Ethics Guidelines for Trustworthy AI.

② 该思想实验的原始提出者乃是福特(Philippa Foot),相关文献有:Philippa Foot,The Problem of Abortion and the Doctrine of the Double Effect,*Oxford Review*,1967,(5).

车前面有一个路人不知何故被绑在轨道上，而电车自己的计算机得出的结论是：现在刹车已经来不及制动以保证电车不撞上这个路人了。此刻，电车的中央控制系统发现电车还有机会将车扳到右边的轨道上去，并由此避免撞上这个路人。然而，就在此刻，电车的传感器又突然发现有五个路人正不知何故被绑在右边的轨道上，而根据电车计算机的计算，电车即使转到了右边的轨道上，它照样来不及制动。所以，电车的自动驾驶系统就只能在“牺牲一人”与“牺牲五人”两个选项之间做出抉择。那么，该系统的设计者应当希望该系统如何做出选择呢？

倘若该设计者本身是边沁的功利主义原则的信徒的话，他就会牢记功利主义的基本教导：“衡量对错的标准是：是否能够让最大数量的人获得最大数量的快乐”[①]，并依据此原则来决定自动驾驶系统究竟该如何在面对电车难题时恰当地决策。按照此思路，似乎该设计者就必须让自动驾驶系统按照下面的程序行事：

功利主义程序一：在系统面对“需要牺牲a条生命”与“需要牺牲b条生命”这两个选项的时候，系统会挑选a与b之间的大者作为保留的对象，并以小者为需要牺牲的对象。

但仔细一想，这样的处理方案未必真正符合功利主义的精神。我们知道，功利主义关心的是如何让全社会的快乐量最大化，而不是让能够活下来的人的数量最大化。我们完全可以设想在某种情况下，这两种最大化并不重叠。譬如，假设在左边轨道上绑着的是某位科学天才，他的创造发明可以让全世界的人获得更多的幸福；而在右边轨道上绑着的都是一些碌碌无为之辈。在这种情况下，摒弃“牺牲五人”的选项并选择“牺牲一人”的选项，就可能产生违背功利主义信条的结果。

由此看来，我们就必须将上面的功利主义程序升级为下面的样子：

功利主义程序二：在系统面对“需要牺牲a条生命”与“需要牺牲b条生命”这两个选项的时候，系统会先计算每条生命所能带来的社会幸福量，然后比较“a条生命带来的幸福总量”与“b条生命带来的幸福总量”的大小，牺牲其中的小者。

然而，执行上述程序却是困难重重的，因为我们很难估计某个人的行为可能给世界带来的幸福总量。譬如，我们可以设想一下爱因斯坦这样的天才在其幼年的夭折会给人类带来的损失与受益：相关的损失或许是我们可能无法发现核能的秘密，而相关的得益则或许是我们可能由此能免于核武器的威胁，但谁又能算清相关的“得”与“失”各自的大小呢？

关于上述问题，边沁本人的解答方案是诉诸对于社会中每个主体的主观快乐程度的心理学度量（用边沁的术语来说，即诉诸所谓的“幸福计算”[felicific calculus]）。现在笔者将在21世纪的技术语境中重述边沁的方案。假设在一个只有五人的极端简易的社会中，一台与特定的大脑

① 这里的“快乐”指愉悦对痛苦的压制。请参看：Jeremy Bentham，“Of the Principle of Utility”，in *An Introduction to* The Principles of Morals and Legislation，T. Payne and Son，eText，1780，p. 1.

窥测设备相连的“边沁牌快乐指数计算仪”将通过如下七个指数采集每个社会成员的快乐指数：(甲)快乐的强度；(乙)快乐的持续时间；(丙)快乐产生的几率大小；(丁)快乐在未来产生的可预期性；(戊)此类快乐产生后，同类快乐紧接着其发生的概率；(己)此类快乐产生后，相反种类的感受(特别是不快乐的感受)不紧随其发生的概率；(庚)此类快乐在社会中被分享的广度。很明显，对于任何的资源分配方案 A 与 B 来说，如果按照上述方式计算得出的与 A 相关的快乐总量超过了与 B 相关的快乐总量，那么，A 方案就会被偏好。将这一思路套用到电车难题上去，我们便可以得出“功利主义程序三”：

功利主义程序三：在系统面对“需要牺牲 a 条生命”(方案 A)与“需要牺牲 b 条生命”(方案 B)这两个选项的时候，系统会先计算面对这两个方案时社会公众所得到的快乐量(相关计算所凭借的数据可以来自相关的心理学或神经科学测试)，然后比较面对方案 A 时公众的快乐量与面对方案 B 时公众的快乐量的大小，选择其中大者所对应的方案予以执行。

程序三显然比程序二更具可操作性，因为在程序三中被计算的乃是心理主体在面对不同的电车难题时解决方案的主观感受(至于如何度量这些感受，实验心理学或神经科学能够提供相应的方法)，而不是相关方案所能够带来的客观功效(关于如何度量这些客观功效，需要大量的观察时间，而很难在当下被立即测度)。然而，即使是这个程序，依然无法处理下述难题：

电车难题的修正版本：假设 R 国与 U 国目前正在交战，而 R 国国民多为极端民族主义分子。在这样的情况下，R 国的心理学被试获知：在电车左边轨道上被绑的是 R 国人，而在其右边轨道上被绑的五个人乃是 U 国人。在这种情况下，按照程序三，对 R 国人来说，看到五个 U 国人被碾死的集体快感会抵消因为看到五个人死而造成的伤感，并使得他们倾向于牺牲五个 U 国人的生命而去保全 R 国人的生命。

这个结论无疑会让很多人感到道德上的不适，而且，即使在 R 国国民中，也很有可能会有少数人对上述结论感到道德上的不适。然而，既然在笔者设想的这个方案中大多数 R 国国民都是极端民族主义者，对上述结论感到不适的少数心理学被试的不快乐感就会被周遭群众的快乐感边缘化，最终成为统计学上的异常数字——尽管任何具有正常道德直觉的人都会觉得这些“少数派”才是更正派的人。这也就暴露出了原始版本的功利主义方案的致命缺陷：他们无法区分君子之乐与小人之乐的质的区别，而试图用某种统一的量纲来处理这两种快乐。基于这种处理方案的社会资源分配方案也难免会因为无法区分快乐之间的内在价值而导致“多数人(小人)对少数人(君子)的暴政”。

原始功利主义的这一缺陷在密尔(John Stuart Mill)的进阶版中得到了一定的克服。密尔所理解的快乐具有更鲜明的利他主义维度，即“独乐乐不如众乐乐”，换言之，每个人都可以通过对社会的奉献而得到更大的快乐。同时，密尔也比边沁更看重某个行为所产生的长远功效，而非其短期利益。因此，这种功利主义学说不太可能在价值上认可那种因集体暴虐行为所导致的快感，

因为这种变态的快感会在更大的时空范围内导致人类的不幸。但从人工智能伦理学的角度来看，密尔的修补方案会带来一个更严重的可计算性问题：既然他已经拉大了考察特定行为之功效的时空范围，那么前文提到的那个难题又出现了——一个有爱因斯坦的世界带给人类的幸福更多，还是一个没有爱因斯坦的世界带给人类的幸福更多呢？从这个角度看，密尔版本的功利主义虽然在直觉上更可爱，但是却是更不可计算的；边沁版本的功利主义虽然在可计算性上貌似略胜一筹（尽管说到底，它实际上依然无法计算不同种类的快乐之间的质的区别），但在直觉上却是令人厌恶的。所以，功利主义无法向人工智能伦理学提供一个能够让我们兼得鱼与熊掌的思想方案。

那么，上文所展示的功利主义对于人工智能伦理学的指导力的匮乏，是不是意味着罗尔斯的社会资源分配理论就更有机会胜出呢？

三、对于罗尔斯的社会资源分配方案的再考察

众所周知，罗尔斯在其名著《正义论》中对于社会资源分配方案的讨论，在学术脉络上与功利主义是彼此分殊的，因为罗尔斯对功利主义的最大批评就是：其社会分配方案由于过于看重大多数人的利益，而无法公正地对待个体的利益。但是在本节中笔者试图论证：在实践中，罗尔斯的社会资源分配方案无法规避功利主义的基本问题，即无法在一个可计算的平台上兑现自身的伦理学承诺。

众所周知，深受康德影响的罗尔斯的资源分配学说之所以在学术史上具有自己独立的生态地位，乃是因为罗尔斯引入了“无知之幕”（the veil of ignorance）的概念（“无知之幕”在此是指参与社会协商的各个主体在某种人为信息屏障的作用下，彼此不知道对方的身份、性别、种族、学历等一切背景信息）。在这一幕布的笼罩下，罗尔斯设想任何行为主体都无法通过对方的背景信息而了解到彼此的力量对比，由此产生的社会契约协商过程也才能变得足够公平、公正。而他本人则试图从这一拟想的社会契约协商过程出发，迂回地得出他自己所心仪的分配正义理论。若用黑格尔的术语来重述罗尔斯的这一思想计划，我们亦不妨这么说：罗尔斯试图将一个正题与一个反题加以综合，然后导出了一个康德式的合题。具体而言，此正题即“伦理学利己主义”（ethical egoism）：根据这种立场，个体的利益总是具有优先性的（顺便说一句，虽然西方哲学史上没有自我标榜为“伦理学利己主义”的大哲学家，但这种立场一向被认为是主流经济学分析的缄默预设）。至于与之对应的反题，则是霍布斯的社会契约论：在这个环节中，原子式的社会个体为了免于“因彼此攻击而最终同归于尽”的恐惧而试图建立一个稳定的社会契约。而在这个正题与反题之后，罗尔斯所引入的合题则是康德式的：换言之，在他看来，只有建立一个能够将任何个体都视为目的（而不仅仅是手段）的社会契约架构，每个人才能保证其基本利益不受侵犯。与之相较，霍布斯本人所主张的“利维坦”式的社会契约架构却不能阻止“利维坦”（即强大的公权力的代称）自

身的肆意妄为，而功利主义式的社会契约架构亦不能阻止个别人的利益被多数人所剥夺。从这个角度来看，任何一个理性的社会契约构造者都不会采纳这两个方案，这是因为，除非他去拥抱康德—罗尔斯主义，否则，他便无法在逻辑上排除自己在未来某一天突然遭受“利维坦”或“多数人暴政”之铁拳的暴击的可能性。

罗尔斯的结论貌似很美好，但站在功利主义的立场上看，该方案的最大问题就是无法解决“资源分配不足”的问题。这也就是说，在能够被分配的物资相对有限而需要物资的社会主体又非常众多的情况下，牺牲某些人的利益其实是无法避免的。对于这一指责，罗尔斯主义者的解决方案是通过引入“机会的均等”来代替对于物质资源的直接平分。譬如，当一个具有罗尔斯主义思想的排指挥官在战场上需要决定让哪位战士先去炸碉堡时，他就会诉诸抽签，以便将每个人生死的几率予以平均化，而为了保障弱小者的利益，他还会主张“对处于最不利位置的人给予最大的帮助”[1](302)，并以此为根据让已经负伤的战士免于抽签；他甚至还会主张“让有官职者也处于公平的机会之下”[1](302)，即让作为排指挥官的自己也参加抽签。这样，即使战斗的结果依然是有伤亡的，但导致这个结果的程序却仍然是公平正义的。

但麻烦的是，在电车难题中，基于下面的理由，这样的程序却是无法被施行的：第一，被绑在铁轨上的利益相关者无法参加投票来决定自己的生死，因为他们没有与司机沟通的信息通道；第二，作为决策者的电车司机无法将自己也视为生死游戏中的利益相关者，因为他并不在轨道上。这也就是说，利益相关者与决策者在时空位置与信息获取权方面的天然不平等，会导致罗尔斯式的机会分配规则在电车难题中成为屠龙之术。

不难想见，如果我们将对于决策所需的信息（注意，这里指的是关于被分配的物资或机会的时空分布的信息，而不是关于别的社会成员的身份背景信息）本身也视为一种资源的话（很显然，对于这种资源的获取本身又是需要别的经济资源与时间资源予以支持的），我们就不难发现罗尔斯方案的致命缺陷：他漏算了这一信息资源在分配游戏中所扮演的角色，而只是缄默地设定这一资源的获取是免费的（但在电车难题里，被绑在轨道上的利益相关者分明是缺乏对于整个困境的全局性信息的）。而且，这种漏算本身是无法在他的理论体系中通过某种技术补丁予以解决的，其理由是：倘若他要再次引入高阶的分配原则以便分配获取这些信息资源的机会的话，为了保持理论的一惯性，他就必须保证这一高阶的分配原则是公平的。但是，为了保证这一高阶的分配原则是公平的，他还要预设与这一高阶原则本身相互匹配的信息是免费的。总之，罗尔斯主义者在此很难不陷入无穷后退。

上面的这个推理也可以从一个相反的角度加以理解：只要我们将罗尔斯漏算的这一因素补上，他的方案就很难不向功利主义的方案塌缩，或至少成为一个带有罗尔斯气味的功利主义方案（比如在功利主义的大方案后面再加一个“额外照顾弱小并限制权力”的补丁）。富有讽刺意味的是，相关的实证心理学研究的确证明了：对于电车难题的“无知之幕”化版本的改写，恰恰更容易

激发心理学被试给出功利主义的应答方式。具体而言，心理学家黄凯伦(Karen Huang)与格林内(Joshua D. Greene)等人针对6000名被试进行了检测，以便调查普通公众对于电车难题等资源分配难题的反应模式[2]。关于电车难题的某个叫“天桥版本”的修正版本，心理学家给出了两个版本：

电车难题的天桥版本：张三在天桥上看着一辆有轨电车在他脚下驶过，且突然发现有轨电车前面有五个人被绑在铁轨上。此时他还估算出电车已经来不及刹车了。这时他又发现另外一个行人也在天桥上往下看这一场景。于是他突然有了主意：将此人推下去，由此制动电车，并救下那五个人(但那个被推下去的人必死)，或者他也可以什么都不做，任凭铁轨上的那五个人去死。

上述版本的“无知之幕”改写版：因为罗尔斯式的“无知之幕”的信息过滤，被试仅仅抽象地知道：有六个人，其中五个人被绑在铁轨上，另外一个人则是相对活动的。他可以被放置到这五个人的前面，起到制动电车的作用，由此救下那五个人。

根据测试的结果，在面对上述第二个版本的电车难题的时候，给出功利主义解决方案(即通过牺牲一个人来救出五个人)的被试数量达到38%，而在面对上述第一个版本的电车难题的时候，给出功利主义解决方案的被试数量却骤降到24%。换言之，恰恰是罗尔斯式的“无知之幕”的介入，导致了功利主义方案在公众中的接受度的明显上升！

读者可能会问：为何在理论层面上与功利主义势不两立的罗尔斯主义会在实践中导致与功利主义的合谋呢？笔者的诊断是，罗尔斯式的“无知之幕”的引入，恰好与功利主义方案的先天缺陷构成了某种共鸣：就像“无知之幕”要消除参与社会契约的个体之间的任何社会差异一样，边沁式的资源分配方案也要故意抹杀处于不同时空位置与社会位置中的个体之间的利益差别，并在此基础上将所有人的利益或快乐都用一致的量纲来加以度量。而这一点也能迅速解释为何心理学被试对于电车难题的天桥版本的反应与其对于该版本的“无知之幕”改写版本的反应彼此不同：在前一个版本中，被试能够清楚地看到那个可能被推下去的行人处在“局外人”的位置中，并因为这种知识而对将一个局外人卷入事端的抉择感到踟蹰，而在后一个版本中，那个行人的“局外人”身份被遮蔽了，这一点反而使得其更容易处在更危险的境遇之中。

此外，需要注意的是，在自觉的理论层面上试图将自己与功利主义拉开距离的罗尔斯其实还给出了一个修补性的分配原则，来适当强化对于个体差异性的照顾，这就是所谓的“差异原则”(the difference principle)。根据该原则，某些社会分配中的不平等是可以被允许的，只要它们的存在能够给社会中的最弱势个体带来更大的利益。譬如，对于某些在生产中的先进个人的物质奖励之所以是能够被允许的，乃是因为这样的奖励能够激发相关个人做出更大的贡献，以便反哺社会，最后惠及弱势群体。但很明显的是，这样的原则依然很难以一种可以被计算的方式得到实施，因为我们很难一般性地了解怎样的物质刺激能够激发先进个人做出更大的创造。具体而言，

有些个体的心理阈值很高，需要更多的物质刺激；有些个体的心理阈值很低，需要的物质刺激很小；有些个体则更看重精神荣誉，而非物质刺激——精神荣誉恰恰是最难被量化的。如果我们要对不同个体的心理倾向做出全面的调研，并据此给出一套更复杂的激励机制的话，我们将由此全面破坏“无知之幕”的预设，并由此使得罗尔斯的方案向亚里士多德式的德性论的方向塌缩。

四、德性论—社群主义的分配方案是否可以被计算化？

与功利主义和自由主义不同，德性论的伦理学与社会框架明确承认了人与人之间的禀赋与社会地位的不同，并依据不同人的内在德性来决定其在社会共同体内所扮演的角色。表面上看来，这样的方案因为引入了个体的多样性而难以被统一的资源分配算法来加以处理，但实际上，与功利主义与自由主义相比，该社会分配方案的可计算性恰恰是最高的。具体而言，从形而上学角度看，德性本身是一个“禀赋”（disposition）性概念，它所牵涉到的是某主体在特定条件下展露出某些具有正面价值的行为的倾向性，比如，“勇敢”这一德性指的就是在需要展现勇敢的外部条件下能够展现出相应行为的倾向。从这个角度看，我们不妨通过对于一个行为者已经展现出来的社会行为的价值进行量化计算，由此计算出其信用积分，并最终将一个人的德性加以量化。当然，这一做法会立即导致自由主义者的抗议，因为这会导致信用积分核定算法的高度垄断化（譬如，会有一个社会权威武断地决定给某类行为以更高的奖励分数，并给某类行为以更低的奖励分数）。但在德性论者看来，这样的武断性本身就存在于德性论所构想的理想社会框架之中，而并非真是其缺陷。

由此，我们甚至不难设想在一个被高度数码化的社群主义社会中人们解决电车难题的方式。在这个社会中，每个人都按照法律的要求随身携带一枚电子手表，手表表面上则会有一个不断变化的二维码来表示其社会信用积分。这样的话，若电车的智能驾驶系统发现了其轨道正前方与支道上都绑有一定数量的行人的话，该系统就会触发自身的传感器去搜集并计算每个行人自身的二维码信息（如果这些行人离开电车太远的话，那么最接近这些行人的摄像头就会将相关的图像输送给电车的中央处理系统）。由此，系统就能计算在主道与支道上所有行人的总积分数（即每个人的积分数与人数的乘积），并择其小者而牺牲之。很明显，该方案要比功利主义的方案更具计算性，因为对于一个人过往的行为数据的搜集难度要远远小于对于一个人的特定行为在未来所带给社会的利益的计算难度。

但即使是这个方案，也会带来四个问题。其一，这个方案预设了每个人的德性表现都能够得到社会信用系统的完整记录。这样的预设显然无法解释像《悲惨世界》中的冉·阿让的表现：按照当时法国资产阶级的法律积分系统，他的德性应当算是差的，但是他却在小说的故事情节的展开过程中充分展现了自己的仁义与善良。倘若在电车难题中，主道上被绑的是被资产阶级政府认可的“优秀”警长沙威，而在支道上被绑缚的乃是五个类似于冉·阿让的通缉犯的话，那么依据

“德性积分比较程序”而运作的电车自动驾驶系统就会选择牺牲五个通缉犯的生命，因为他们的德性总积分可能都抵不上沙威警长一个人的积分。但很显然的是，这样的结论是违背德性论自身的直觉的，因为我们都知道这五个所谓的通缉犯都是错误的社会处罚体系的牺牲品。

其二，即使我们搁置在上一个自然段中所展现的“社会信用系统无法完美记录个体的真实德性表现”这一问题，基于“德性积分比较程序”的社会资源分配方案还是会错误地做出如下预设：具有某种德性表现的某一个体还会在未来继续给出类似的德性表现。很显然，这样的预设无法解释三国时期东吴的“周处现象”：鄱阳太守周鲂之子周处年少时纵情肆欲、为祸乡里，后来浪子回头、改过自新，功业更胜其父，留下“周处除三害”的传说。很明显，在周处生命的某个转折点，他的德性表现发生了重大的改变。现在我们就可以利用这种价值观转变的案例构造出一个新的电车难题场景，以便展露出基于“德性积分比较程序”的社会资源分配方案的短处：

电车难题的周处版本：假设浪子周处迄今为止所获得的德性信用积分都是很低的，但在一个时辰之前，他突然醒悟，决定重新做人。这时候，他在天桥上看着一辆有轨电车在他脚下驶过，且突然发现有轨电车前面有五个朝廷的高级官员被绑在铁道上（在这里我们默认这五位高官都有很高的德性积分）。此时他还估算出，电车已经来不及刹车了。于是他有了主意：自己跳下去，由此制动电车，并救下那五个人。但这样一来，他自己就肯定活不成了。

为何说周处版本的电车难题会暴露出“德性积分比较程序”的不可计算性呢？这是因为该版本展现了基于德性积分的计算结果与德性论所支持的道德直觉之间的重大分歧。按照德性论者的标准——道德直觉，一个愿意舍己救人的人显然是具有很高德性的，所以，周处的自我牺牲行为值得最高程度上的道德褒奖。与之相较，按照德性积分的计算结果，周处本就活该去用自己的肉身挡住电车，因为他的德性积分本来就很低。但这样一来，一个被动地被人推下天桥的周处与一个主动跳下天桥的周处之间的德性差别也就被抹平了，而这一抹平效应本身就是违背德性论者的标准——道德直觉的。当然，基于德性积分的计算系统也能在周处牺牲后再去追加他的道德积分，但这一具有滞后性的做法还是无法在周处跳桥的那一刹那区分出周处主动跳与周处被推下去之间的差别。抑或一台灵敏性极强的基于德性积分的计算系统能够在周处跳下去的3毫秒内迅速改变他的道德积分，由此勉强区分周处主动跳与周处被推下去之间的差别。但即使这样做，该系统依然会陷入某种尴尬：假设周处这时候得到的额外德性积分是如此之高，以至此分数已经超过了被绑在铁轨上的五个高官的德性总和，那么，按照系统的内在程序，周处是不应当跳桥的。这就导致了一个有趣的悖论：如果周处跳了，他就不应该跳；如果他不跳，他就应该跳。很显然，这样的悖论会导致系统自身的宕机。

其三，基于德性积分的计算系统还面临着另外一个问题：按照标准的德性论—社群主义模型，能够给出德性评判的社会权威来自不同的社会等级（包括家庭、宗族、社区、教区、俱乐部、学校、行业共同体、方言共同体、地方行政单位、国家等），由此呈现出丰富的社会生态体系。且不提

对如此多的德性积分积累体系进行算法化是否可行(至少从常识上看,对家庭内部的德性积分进行面向机器的算法化处理乃是相当疯狂的),即使这种全面的算法化是可行的,不同的社会亚系统所具有的不同德性评价方式所带来的彼此兼容性问题也会变得非常尖锐(譬如,在电影《闻香识女人》中,美国中学生查理在其人生的某个阶段,就必须面对"背叛同学,获得校长好评"与"得罪校长,保守同学秘密"这两个彼此冲突的选项。而这两个选项中的任何一个,都受到了相关的社会共同体的德性积分系统的支持)。此外,由于对于人类社会中既有的非算法化的德性积分系统进行全面的面向机器的算法化作业肯定需要消耗大量的社会资源,所以,我们不难设想:只有那些得到巨大权力与资本支持的超级社会团体才能够将自己的意志灌输到上述算法化工程之中,由此德性论原本所看重的社会生态的多样性也会遭到破坏。

其四,古典德性论的代表人物亚里士多德所最看重的德性具备"实践智慧"(*phronesis*)——这是一种在恰当的场合中将善良意图以最合适的方式予以贯彻的能力。众所周知,亚里士多德特别强调个体在社会中反复历练以获取这种实践智慧的重要性,并强调了经由这种智慧所做出的道德决策对于特定语境的敏感性。虽然亚里士多德本人并没有明确提到电车难题,但是从他关于实践智慧的一般性阐述原则中我们不难推出:从他的立场看,应当不存在关于电车难题的某种一般性解法。毋宁说,试图解决此类问题的当事人必须身处这样的两难问题处境:切身感受该处境的一切细节,并根据他自己的人生经验来做出临机决策,而不能根据对于该处境的某种抽象的纸面报告做出道德判断。这一处事原则显然就拒斥了通过某种抽象的德性算法来决定牺牲谁、保全谁的可能性,因为这种算法会彻底压缩当事人运用自身的实践智慧的空间,并因此成为一种反实践智慧的算法。

从以上四个方面的分析来看,德性论的算法化指数虽然比边沁的功利主义与罗尔斯的自由主义方案略高,但其实也只是"五十步笑百步"罢了。换言之,几乎所有主流的规范伦理学进路都很难通过某种算法化工程而成为人工智能伦理学的指导方案。

五、深入讨论

几乎所有主流的规范伦理学进路都难以被真正算法化这一结论,显然向我们暗示了某些更重要的真理。从更抽象的角度看,数据化的本质是一种彻底的第三人称视角的贯彻:一切都可以由此被公开地加以计算与处理,如同天文学家对于行星轨道的计算。同时,数据化也意味着某种可复制性——任何数据都能够被复制到别的信息载体中,成为其运行的基础,由此抹杀不同的数据载体之间的区别。与之相较,人类社会的运作却是建立在公开性与隐私性之间的平衡之上的:一方面,人类作为社会动物的本质的确需要人类意图与行为的恰当公开性,但在另一方面,人类共同体的多层次性却要求在各个层次上建立起某些信息屏障,譬如,有些关于个体的内部信息不必告诉别的家庭成员,一些家庭的内部事务不必告知社区,一些公司内部的事务不必告知社会,

等等。很显然，自由主义所看重的个体隐私与社群主义所看重的社会共同体的多样性都是建立在这种微妙的平衡之上的。因此，一种彻底消灭隐私并将一切诉诸算法的社会，将意味着人类各种传统价值的全面崩溃。

这样的结论无疑会给人工智能伦理学在分配正义领域的全面运用蒙上重重的阴影。不过，这一结论并不意味着某种辅助意义上的算法处理不能在分配领域发挥特定的作用（实际上，几乎没有人否认计算机在资源分配过程中所起到的工具性作用），甚至亦并不意味着我们不能在“通用人工智能”的大背景中使得一个独立运作的人工智能体最终具备接近人类水平的资源分配能力。但需要指出的是，即使这样的智能体能够获得亚里士多德式的实践智慧能力，它也不会机械地将某种普遍化的伦理算法套用到各种不同的分配案例上去，而会自主进入不同的场景，像人一样进行“现场感知”。不过，要实现这个目标，这样的人工智能体就得具备全向的感知、记忆、推理、共情能力，并在此基础上构筑出自身的伦理推理能力。很显然，要实现这个目标，我们还需要海量的理论预研。因此，仅仅就目前的技术状态而言，在现有的人工智能技术与现有的关于分配的规范性理论之间进行直接的嫁接，乃是非常鲁莽的。

不过，即使我们能在“通用人工智能”的大背景中重启人工智能伦理学的研究，我们也需要看到，人工智能体在资源分配问题上体现出来的“实践智慧”只能是对于人类特定类型的实践智慧的一种迁移。很明显，不同的文化共同体都会给出能够体现自身文化特色的不同的实践智慧落地方案，譬如，在文化A中，做A会被认为是富有实践智慧的，而在文化B中，这么做却会被认为是缺乏实践智慧的。在这种情况下，在不同的文化共同体中依照不同的伦理习惯而被构造出来的不同人工智能体，也只会以算法化的形式将关于实践智慧的文化差异予以固化，由此使得某种适用于全球各文化的关于分配正义的算法再一次成为空中楼阁。这或许对于本来就看重共同体生长的特殊历史的社群主义者来说是一个利好消息，但对于功利主义者与义务论者来说则肯定不是，因为他们的梦想恰恰是建立一个横跨所有文化历史差异的普遍性资源分配方案。

参考文献

[1] John Rawls, *A Theory of Justice*, The Belknap Press of Harvard University Press, 1971.

[2] Karen Huang, Joshua D. Greene, Max Bazerman, Veil-of-ignorance Reasoning Favors the Greater Good, *Proceedings of the National Academy of Sciences of the United States of America*, 2019, 116(48).

责任编辑：陈　菊