武漢大學學報(哲学社会科学版)
WUHAN UNIVERSITY JOURNAL (Social Science)
第70卷 第6期 2017年11月
Vol. 70. No. 6 Nov. 2017. 049～059
DOI:10.14086/j.cnki.wujss.2017.06.005

儒家德性伦理学、神经计算与认知隐喻

徐英瑾

摘　要：一般而言，“德性伦理学”就是指通过道德主体的行为倾向，来对其道德输出进行评价的伦理学路数。儒家的道德学说在很大程度上可以被归类为德性伦理学，并完全可以在一个自然主义框架中得到“祛魅化”处理。这种处理的第一步，便是按照扎格泽布斯基的“行为者德性论”与思汪敦的“击靶德性论”的话语方式，厘清儒家德性论的基本框架；而在第二步中，我们不妨深入探讨经由某种可计算的平台来进一步刻画“德性熏养”机制的技术可能性，并由此彻底地将儒家的传统话语方式顺化为当代认知科学的话语方式。文章认为，丘奇兰德的神经计算模型对于德性熏养机制的刻画，将会不可避免地遭遇到“描述层次崩塌”的难题，因此并不适合成为刻画儒家德性伦理学的恰当工具；而一种更符合儒家精神的技术刻画模型，将诉诸种种基于“儒家道德样板库”的隐喻投射机制，使得此类刻画得以“可计算化”的技术平台，则是“非公理化推理系统”。

关键词：儒家；德性伦理学；击靶德性论；神经计算；隐喻
中图分类号：B222；B84　　**文献标识码**：A　　**文章编号**：1672-7320(2017)06-0049-11

一、导论：再谈德性伦理学与儒家哲学的结合

众所周知，如何在当代思想的语境中激活古老的儒家学说的意义，乃是一个备受当下汉语思想界关注的重大课题。然而，在汉语语境中不太被提及的是，在英语世界中，已经出现了一个运用现代科学与分析哲学的资源来重新解释儒家思想资源的新倾向。譬如，加拿大认知科学哲学家兼儒家学者森舸澜(Edward *Slingerland*)便基于认知神经科学方面的证据，指出：重视道德体验——而不是明述化的道德推理——的儒家伦理学学说便是一种能够得到现代认知神经科学证据印证的理论，因为根据神经科学的研究，人类所作出的很多重要的道德判断，的确是在负责高级思维的脑区未被激发的情况下，以“不费力”的方式给出的(Slingerland，2010：247-286)。很明显，森舸澜所提到的那些能够给出“不费力”的道德判断的道德主体，肯定是具有一定的“德性”的(否则道德判断的准确性就难以被担保)——因此，他的这一叙述方式本身就包含着与德性伦理学的某种关联。而近几年在这一方向上做出较多用力的，乃是台湾地区东吴大学的米建国教授。他与美国知识论界的重磅级哲学家索萨(Ernst Sosa)(Sosa，2015：325-330)合作，已经开启了一个用英美德性论的思想资源重解儒家思想的新研究方向，引发了两岸学界的一定关注(Mi，Slote & Sosa，2015)。

不过，米建国教授与索萨教授更为关心的是“儒式德性知识论”的构建，而不是本文所关心的德性伦理学问题，因此，从伦理学角度将儒家的学说予以“德性化”的理论空间依然很大。然而，考虑到儒家学说与德性伦理学各自又都具有不少分支流派，过于笼统地谈论结合二者的可能性，或许并不能真正有助于将对于问题的讨论引向深入。为了避免这种笼统性，本文将采取如下的研究路线图：在第一节中，笔者将讨论目下英美学界阐发德性

伦理学的四大路数,并就这些路数与儒家哲学中特定资源的对应关系进行提点。在这个环节中,笔者将特别关注"击靶德性论"与儒家纪传体史书中的人物评价传统之间的关系。在第二节中,笔者将进一步讨论在计算机建模的环境下重构儒式人物德性品评传统的可能性。而在第三节中,笔者将转而讨论基于"儒家德性样板库"的隐喻性投射来熏养人工系统之"德性"的新技术进路,以便提供一个(比现有的神经计算构架更好的)用以理解儒式人物品评传统的语义模型。

本文所进行的研究的元哲学预设是自然主义的,既认为儒家德性伦理学是可以在一个诉诸当代自然科学的新话语框架中得到"祛魅化"处理的。正是受到这种元哲学观点的激励,笔者在选取与儒家思想相关的历史资源时,除了会适当地关注以《四书》为代表的儒家传统经典之外,也会特别留意受到儒家思想影响的历史学家对于真实历史人物的德性评价资料——因为后一类资料显然比前者更容易被一个自然主义的话语框架所吸纳。同时,这样的处理方式,也在某种程度上迎合了马克思与恩格斯在《德意志意识形态》中表达的唯物史观的基本论点:意识形态(包括儒家经学传统)是没有自己的独立历史的,除非这些意识形态能够被兑现为对真实历史中真实人物真实活动的记载。

二、四种德性论及其与儒家学说中的关系

在西方伦理学文献中,"德性"(virtue)这个词大略上指的是一个道德主体在特定种类的外部条件的刺激下给出特定种类的道德输出(如道德欲望、道德感受、道德行为)的倾向(dispositions),而且这里所说的道德输出肯定是具有"善良"、"美好"这样的正面价值的。这里需要注意的是,从形而上学的角度看,"倾向"这个词是具有对于"反事实条件"的支持力的——譬如,一个具有"勇敢倾向"或"勇敢德性"的人,即使在没有机会展示其勇敢行为的环境下,依然是"勇敢的",因为他可以在"出现危险"这一反事实条件被满足的情况下向大家展现出勇敢的行为。不难想见,"德性"的这种特点,可以使得拥有相关"德性"的主体的行为模式在观察者那里得到稳定的预期,并由此使得这些被评价者得到社群的信任。而所谓"德性伦理学"(virtue ethics),也就是对所有将"德性"这个概念视为伦理学基本概念的伦理学立场的总称。

在西洋规范伦理学的谱系中,德性伦理学的对立立场主要有道德义务论与道德后果论。非常粗略地说,义务论者关心的道德行为是否基于应然性的道德规范,而后果论者关心的是道德行为是否能够带来功利的效果。至于德性论与义务论以及后果论之间的本质性差异就在于:德性论关心的乃是给出道德行为的人或者团体,而后二者关心的则是道德行为本身。因此,若用史学史的术语来打比方说,德性论者天然就偏好于"纪传体"的世界描述方式(因为纪传体的写法就是"以人带事的"),而义务论与后果论者会更偏好于"编年体"的世界描述方式(因为纪传体的写法就是"以事带人的")。

在笔者看来,尽管我们无法以绝然的方式断定儒家伦理学的理路与西方义务论/后果论毫无瓜葛,至少我们可以在这一理论与西方德性论之间找到更多的类似之处。非常粗略地说,德性论的一个重要理论优势是可以通过比较便捷的方式确定一个行为在道德上的可接受性——也就说,一个行为是否正确,主要取决于行为者的德性,而不是行为自身的道德根据与历史后果。这就在很大程度上规避了义务论者与后果论者都难以解决的两个问题:(甲)在很多场合下,对于行为的道德根据的追索会遭遇到彼此冲突的义务论规范;(乙)在很多场合下,对于一个特定行为的真正后果是难以被预料到的。而诉诸"德性"的道德理论,则可以籍由行为者处理复杂道德处境(特别是那些包含着彼此冲突的道德要求的道德情境)的卓越能力以及对于自身行为的后果的预见力,而使得对于上述问题(甲)与(乙)的明述化解答变得不那么必要了(注意:这里所说的这些能力也都是"倾向性"概念)。很显然,这一理路与孔子在《论语·为政》中的著名表达——"……四十而不惑,五十而知天命,六十而耳顺,七十而从心所欲,不逾矩"——在精神上是彼此暗合的,因为孔子的这一评论主要就是针对有德性者处理问题的一般能力倾向的熏养过程而言的,而不是针对某个具体行为或事件而发的。

不过,有鉴于德性伦理学目前也已经发展出了不同的学术分支,为了将问题的讨论引向深入,我们还需要对这些分支与儒家思想资源的具体对应关系进行更细致的耙梳。在进行相关讨论时,笔者参考

了胡斯特浩思(Rosalind Hursthouse)为在英语世界具有权威地位的《斯坦福哲学百科全书》中的“德性伦理学”词条①所给出的知识梳理框架。

第一种要被顾及的德性论品种乃是“柏拉图式德性论”(Platonic virtue ethics)。此论预设柏拉图式的“善”的理念有一种独立于心灵的存在,并认为:德性熏养的要点就在于,我们要对我们所遭遇到的万事万物所蕴藏的“善”进行凝思,思考其自身的利益,琢磨其内在的品性,而不能将目光聚焦于与世界割裂的“小我”(Chappell,2014:300)。由此,按照此论,“德性”的实质便是对于一系列外在价值对象的领悟能力。另外,还有一些基督教神学背景更明显的柏拉图式德性论者,将外部“善”的理念的根据视为人格化的上帝,并将个体的德性熏养过程,视为其对于上帝在“爱”这个维度上的模仿度不断上升的过程(Adams,1999:36)。从中国文化的立场上看,虽然“柏拉图式德性论”所涉及的柏拉图主义与基督教神学的思想资源均与儒家资源具有很大的异质性,但这种借助某种独立于心灵的外部资源提升内部德性的思路,对于儒家的某些分支学派来说却并非隔膜之物。譬如,朱熹对于《大学》中“格物致知”一语的解释——“一书不读,则阙了一书道理;一事不穷,则阙了一事道理”(《朱子语类》卷15)——在思路上便与柏拉图式德性论“经由凝思客观之善而提高自身德性”的理路暗合。

第二种要被谈及的德性论品种乃是“幸福式德性论”(Eudaimonist virtue ethics)。此论的关键词是“幸福”,其希腊文原文是“εὐδαιμονία”,其拉丁转写形式是“Eudaimonia”,而“幸福”则是对于这个古词的某种非常勉强的今译。“Eudaimonia”兼指肉体与精神方面的满足(且略偏重于精神满足),其形容词形式“Eudaimonist”主要用来修饰“生命”、“生活”这样的名词,因此,“幸福式德性论”的“世俗色彩”要浓于柏拉图式德性论。依据幸福式德性论,“幸福”与“德性”的联系在于:(甲)二者都是道德概念;(乙)在不少哲学家看来,良好的德性是通向幸福的必要条件,尽管关于这一必要条件是否同时是充分条件,则见仁见智。需要指出的是,这种将“福”、“德”并提,并借此刻画“德性”的思路,自苏格拉底、柏拉图、亚里士多德以降便一直是西方古典德性伦理学的“正路”,直到二十世纪德性论复兴后,其地位才被“行为者德性论”与“击靶德性论”所取代(详后)。不过,笔者认为,要在“幸福式德性论”这一维度上展开儒家思想与德性论的有效对话,恐怕并不容易,因为我们很难在儒家的词汇表里面找到“Eudaimonia”的严格对应者。相比较而言,相对接近“幸福式德性论”之精神的儒家案例乃是“颜回乐道”,以及《孟子·梁惠王下》提到的“独乐乐不如众乐乐”。然而,儒家并没有将这里所提到的“乐”提炼为一个地位堪比“Eudaimonia”的哲学范畴予以全面阐发——相反,正如杨泽波先生所指出的,在更多的案例中,儒家更倾向于从“偶然性”的角度审视对于“福”与“德”之间的关系,即认为:即使在坚持德性的前提下没有得到福报,君子也应当将此当成一种“命运”来坦然接受(杨泽波,2010)。

第三种要被讨论的德性论品种乃是“行为者德性论”(agent-based virtue ethics)。按照此论,道德规范的根基就在于道德行为者(moral agent)自身的品性与行为倾向与动机——譬如,一个行为到底对不对,取决于道德行为者具有怎么样的道德动机,或基于其哪方面的品性——如果是出于其善的动机,或者是其人格中美善的一面,这样的行为便是好的(Slote,2001:14)。或者说,一个错误的行为,就是一个既有实践智慧(Phronesis)的人通常不会做的(Zagzebski,2004:160)。这里需要特别指出的是,按照美国女哲学家扎格泽布斯基(Linda Zagzebski)的看法,我们对于“那些动机是好的,哪些是不好的”这一点的判断本身又是基于对特定道德榜样(exemplar)的回应(Zagzebski,2010:41-57)。这样的道德榜样未必是某个特定的人,而可能是我们在历史上所遭遇到的众多道德高尚的人的某些共通点所汇聚成的价值网络——而个体对于此类价值网络的浸淫,则可以帮助前者在特定的道德情境中以恰当的方式模仿先贤,给出精准的道德判断与合适的道德行为。

相比较前两类德性论而言,笔者认为行为者德性论与儒家学说的关系更为密切。大致而言,柏拉图

① Rosalind Hursthouse. “Virtue Ethics”. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (eds.). URL=https://plato.stanford.edu/entries/ethics-virtue/. 2017-07-30.

式德性论至多只能与较晚出现的程朱理学发生积极的联系,而与原始儒家的关系比较疏远;至于幸福式德性论对于"Eudaimonia"的聚焦,也在儒家资源中缺乏足够明显的理论对应物。而在儒家的传统资源库中,能够印证"行为者德性论"的描述则要丰富得多。前文已经引用的孔子在《论语·为政》中对于个体德性的培养过程的自传性描述,几乎是以一种惊人的精确性预报了扎格泽布斯基论点的要旨:德性的熏陶需要丰富的人生经历来积累足够多的道德样板,以便使得正确的道德动机的发端能够获得足够好的模拟对象。甚至孔子本人对于《春秋》的编纂所具有的意义,也可以在某种广义的行为者德性论的框架中重新加以解释:具体而言,《春秋》本身就是对于历史上所积累的正、反两方面的道德案例与德性展示的典型化处理,其目的便是为了激发阅读者正确的道德直觉,培养其自身的德性——所以古人才有"孔子成《春秋》,而乱臣贼子惧"(《孟子·藤文公下》)这一说法。由此推演下去,从司马迁开始的中国纪传体史书编纂实践,则可以被视为对于孔子原始"道德样板数据库"的不断的"扩容"工程。这一经由历史上既有的道德样板衡量当下人物德性的传统是如此之强大,以至于在清末的徐继畬对于乔治·华盛顿"起事勇于(陈)胜、(吴)广,割据雄于曹(操)、刘(备)"(《瀛寰志略》)的评价中,我们都能隐隐看到此类儒家思维模式的影响。

不过,从学理上看,若真按照儒家史学传统的评价实践去衡量,扎格泽布斯基的行为者德性论依然还不够精致,因为此论更为关心的是有德者在特定环境下所会给出的道德动机,而在儒家的人物品评实践中,特定行为所导致的结果也会成为德性评估的参照指标。以范晔在《后汉书·虞傅盖臧列传第四十八》中对于东汉末年的人物臧洪(160 — 195 年)的评价为例。在这则案例中,臧洪因为袁绍没有发兵去救其好友张超而在东武阳起兵反袁,后因城困粮尽,兵败被杀。范晔一方面以"洪怀偏节,力屈志扬"这样的评价来表扬臧洪,又批评他不懂"忿悁之师,兵家所忌"的军事常识,甚至嘲讽他想学当年申包胥搬秦兵救楚而不得①。而在笔者看来,若范晔能够学会英美伦理学的"行话"的话,他或许也会这样来重新组织他的评价修辞:臧洪反袁的动机固然出于为友人复仇的良好动机(我们知道,流行于汉代的《春秋公羊学》本身是认可一定条件下的复仇行为的),但他却没有以恰当的方式实现相关目标的其它德性——比如军事上的审慎与政治上的判断力。因此,至少臧洪的悲剧足以说明,过分看重"良好动机"的行动者德性论并不足以构成对行为规范的完整说明。

不过,"行动者德性论无力说明臧洪案例"这一点,并不意味着德性论整体框架的失败,也并不意味着后果论理论的某种胜出(尽管从表面上看来,范晔对臧洪的批评,多少是建立在对于其行为的消极后果的观察之上的)——因为德性论完全可以通过升级自己的理论框架,而将某些后果论的因素包容于自身。这就引出了本节所讨论的最后一种德性论类型:击靶德性论(target-centered virtue ethics),提出人为美国奥克兰大学的女哲学家思汪敦 (Swanton,2003)。与前几种德性论不同,击靶德性论者并不喜欢抽象地谈论"德性"概念,更不喜欢引入"幸福"、"善"这些更抽象的概念,而是致力于将"德性"兑换成平时我们所经常用到的"德性"名目,而"勇敢"、"诚实",等等。与重视动机反省机制的行为者德性论尤其不同的是,击靶德性论特别看重德性价值的实现(故此才有"击靶"一说)。譬如,"勇气"这一德目之标靶如果说是"克制愤怒、正面危险"的话,那么,只有真正做到这一点,相关德性才能够得到实现,"击靶"活动也才算真正完成。正是基于这种观察,按照击靶德性论,一个行为是否在伦理上可被赋予正面价值,将取决于该行为是否击中了自己的道德标靶,而不像行为者伦理学家所说的那样,主要取决于该行为是否在意向中瞄准了相关的道德标靶。此外,由于"德性"一开始就是作为一个复数概念进入击靶德性论的词汇,因此,如何对多重德性同时提出的"击靶要求"进行全盘考量,也一直是击靶德性论者所关心的问题。在他们看来,在特定的环境下由于某项更重要的"击靶要求"而放弃某些可能会与之发生冲突的次要"击靶要求",不仅在道德上是可以被允许的,甚至在实践中也是不可避免的,因为个体的时间与资源无法同时满足那么多的"击靶要求"。因此,一个行为的德性属性,最终也将取决于上述这种通盘考虑的

①这里的历史背景是:臧洪曾指望公孙瓒、黑山军、吕布能在侧后袭袁以救东武阳,不料其希望却统统落空。

恰当性，尤其是其与特定语境的适切性。

不难看出，击靶德性论在很多方面都与儒家伦理学高度契合。第一，作为复数概念的“德性”对于儒家来说毫不陌生，“四维”（礼、义、廉、耻）、“五德”（智、信、仁、勇、严，或：温、良、恭、俭、让）、“八德”（孝、悌、忠、信、礼、义、廉、耻）这样的说法都是广为中华文化圈所知的。第二，在范晔对于臧洪的评价中我们已经看到了，儒家对于德性的实现——或“击靶”——有着独特的兴趣。此外，范晔之前的叔孙豹早就将“立功、立德、立言”视为“三不朽”（《左传·襄公二十四年》）；而更往后看，范晔之后的孔颖达又在《春秋左传正义》中将“立功”进一步界定为“立功谓拯厄除难，功济于时”——可见对于事功之成败的关注，无论在叔孙、范、孔那里，还是在思汪敦那里，都别无二致。其三，儒家对于特定语境中问题解决资源的有限以及不同德目之间的内在冲突，是有着清醒认识的，所以儒家才特别关注“如何在这些价值目标之间找到恰当的平衡点”这一问题。《论语》本身就提供了不少这样的平衡式案例，譬如：对于“照顾父母”与“切实的出游需求”这对矛盾来说，“游必有方”就是这样的平衡点（《里仁》）；而对于在动荡政治语境中“全身”与“善道”的平衡问题，孔子开出的折衷性药方则是“危邦不入，乱邦不居”，等到政局安定后再“有道则见”（《泰伯》），等等。此外，也正是按照类似的标准，范晔才在《后汉书》中对臧洪反袁行为的不审慎作出了批评——或改用击靶德性论的话语来说，范晔笔下的臧洪，缺失了“对多种复杂因素进行全盘统筹”这一与“击靶”密切相关的德性。

从本节的分析来看，扎格泽布斯基（以下简称“扎”）的行为者德性论也好，思汪敦（以下简称“思”）的击靶德性论也罢，它们要么诉诸历史上积累的道德样板资源对于个体德性的熏养机制，要么诉诸有德性的个体的内部道德决策机制——总之，它们既没有像柏拉图式德性论那样求助于超自然概念，也没有像幸福式德性论那样求助于“Eudaimonia”之类的语义模糊的概念。这些特征，无疑使得扎、思之论具有了非常明显的自然主义面相，并由此使得它们有资格进一步成为备选元叙述框架，使得“儒家资源之自然主义化”这一议题得以展开。

然而，一个彻底的自然主义者，将不会仅仅满足于将德性论中某些更具自然主义意味的分支与儒家伦理学互相捆绑。一种更彻底的自然主义方案，将通过引入某些经验科学（而不是现代伦理学）资源，以便对儒家伦理学施加更为深入的“祛魅化”操作。这也正是本文余下两节所要尝试着去做的。

三、从神经计算模型看德性熏养

非常粗略地说，现代自然科学的特点便是表述诉诸量化手段，结果可由经验所验证。那么，怎么使得儒家的德性论也成为这样一种科学化的“心性”理论呢？一种比较容易想到的思路，便是像前文所提到的森舸澜那样，寻找利用神经科学的证据来印证儒式道德理论的可能性。但有鉴于神经科学的描述的层次远远低于德性描述的层次，这样的进路是否会导致解释中“层次不匹配”的问题①，笔者是有所担忧的。而为了纾解这一问题，一个很容易想到的方案，便是提高原来的神经科学描述的层次，由此使得一种关于德性的高层次理论能够附着于其上。

这种比神经科学更高，却依然与其有关系的描述层次，就是“神经计算”的层次——在这个描述层次上，人工智能专家对于人类神经网络活动的方式进行适当的数学抽象与模型化，并凭据这一模型来对很多科学假设进行验证。与之相关的技术路径，则在人工智能文献中一般被称为“人工神经元网络”（artificial neural network）或“联接主义”（connectionism）。

非常粗略地说，神经元网络技术的实质，就是利用统计学的方法，在某个层面模拟人脑神经元网络的工作方式，设置多层彼此勾联成网络的计算单位（如输入层—隐藏单元层—输出层等）。由此，全网便可以以类似于“自然神经元间电脉冲传递，导致后续神经元触发”的方式，逐层对输入材料进行信息加工，最终输出某种带有更高层面的语义属性的计算结果。至于这样的计算结果是否符合人类用户的需

①在认知科学的语境中，“层次”指的是对于特定科学描述的宏观／微观程度的描述。层次越高就越宏观，层次越低就越微观。

要,则取决于人类编程员如何用训练样本去调整既有网络各个计算单位之间的权重(参见图 1)。一般而言,隐藏层计算单元只要受过适当的训练,就能够初步将输入层计算单元递送而来的"材料"归类为某个较为抽象的范畴,而所有的这些抽象范畴之间的语义关系,则可以通过某种记录隐藏层计算单元之触发模式的所谓"矢量空间",而得到一种立体几何学的表征。

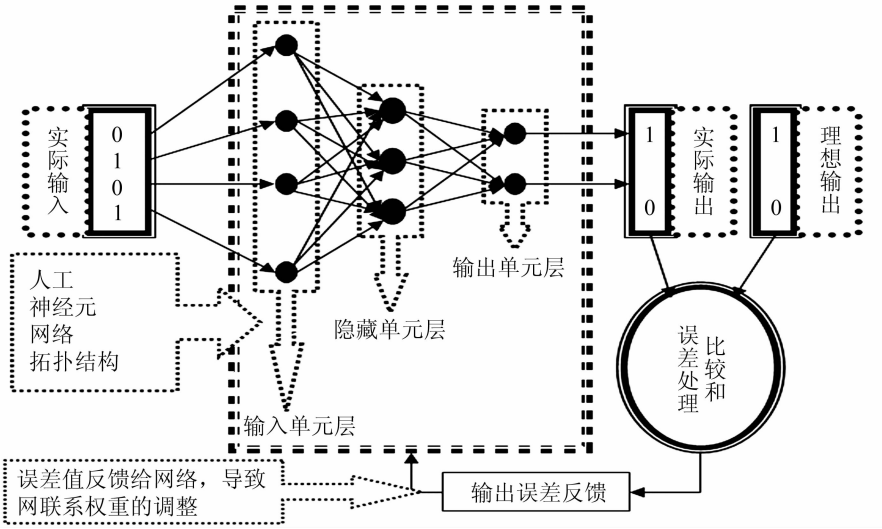

图 1　一个被高度简化的人工神经元网络结构模型

基于上述基本技术思路,"神经哲学"的倡导者、美国哲学家丘奇兰德(Paul Churchland)便设想了在某种经过精心调试的人工神经元网络的平台上完成"德性"训练的可能(Churchland,2007:37-63)。他的大致想法是:如果我们能将关于人类行为的某些基本底层描述全部数码化并"喂入"一个神经元网络的话,那么,通过调整网络节点之间的信息传输权重,我们就能够使得网络中的隐藏层形成一个关于道德价值词分布的"矢量空间"(图 2)(Churchland,2007:43)。而在这样的矢量空间中,每一个离散的点都表示网络对于特定某种行为样板的典型性表现形式——比如,"撒谎"[lying]这个点就是与撒谎有关的各种输入在得到特定的处理后,隐藏层所应该在矢量空间中激发的位置。而这些激发位置在矢量空间中构成的几何体,则形象地表示了一个具有特定价值观的行为者所尊奉的价值体系的内部结构。

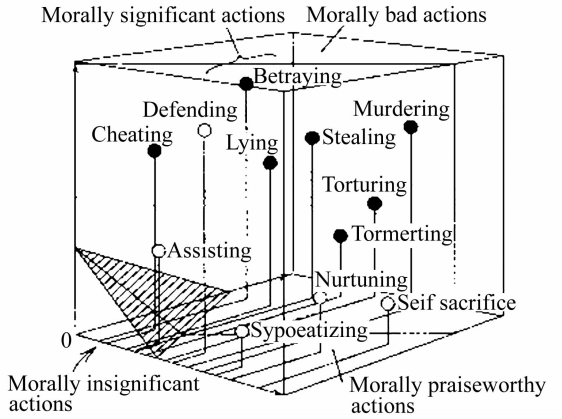

图 2　丘奇兰德所设想的神经网络模型经过训练后形成的关于道德识别的矢量空间

注:图中所有实心小圈表示在道德上值得谴责的行为类型在矢量空间中的典型位置,所有空心小圈表示所有在道德上值得表扬的行为类型在上述空间中的典型位置。代表善行的小圈所在的亚空间与代表恶行的小圈所代表的亚空间又在纵向上将整个矢量空间剖分为二——此外,该矢量空间的左下角被斜切出来的四面体则代表"缺乏道德意义"的中性亚空间。

虽然丘奇兰德没有提到儒家思想资源与他的技术建模之间的关系，但是对于此建模对于一般意义上的德性伦理学的说明意义，他却有着自觉的意识。依据他的理路，在神经计算的话语框架中，所谓“获得一种德性”，就是通过特定的数据训练而能够做到以下两点：(1)在关于隐藏层触发模式的矢量空间中，形成一个在几何学特征与拓扑学特征上均符合社会规范要求的“价值几何体”(或说得通俗一点，知道“撒谎”与“诚实”之间的概念差距要大于“撒谎”与“保持沉默”之间的差距)；(2)在遇到新的行为输入的前提下，能够将这些输入正确地映射到前述“价值几何体”的正确位置上去(或说得通俗一点，知道哪些行为算“撒谎”)。而一个网络一旦满足了此类训练要求，我们就可以信赖其能够在未来遇到新的行为刺激时继续产生符合社会规范之期望的道德输出，因为它已经获得了特定的“德性”(Churchland，2007：43-47)。

不得不承认，至少从表面上来看，丘奇兰德的德性伦理学的刻画思路的确与儒家的思路有些相合：二者都试图在“前命题推理”的层面上处理德性的“熏养”问题；而且，二者都承认这种“熏养”需要外部的社会权威的介入(说得具体一点，在人工神经元网络中，系统初始输出与理想输出的比对，就是“社会权威意志”对于人造系统的一种介入方式)。而且，丘奇兰德对于“价值几何体”中的“价值端点”的复多性的强调，也与击靶伦理学的类似想法相似。然而，出于如下理由，笔者依然认为：要将丘奇兰德的相关技术设想完全落实到对于儒家资源的重建上去，我们肯定会遭遇到重大的困难——而且这些困难不仅是技术性的，而且是带有哲学面相的。

为何这么说呢？我们知道，在当前人工智能的实践中，人工神经元网络经常被用以执行“模式识别”的任务，譬如从海量的照片或视频中辨认出特定人员的面孔。而这类任务的基本特点是：其输入与输出之间存在着一种“描述层次逐步增高”的过程，而这种过程也与神经元网络自身的金字塔式构建形成了某种对应。而在道德判断的案例中，我们却很难找到这种类似的层次性。这又是因为：

第一，从儒家的历史叙述传统来看，我们很难找到不包含高层面价值评价的低层次物理描述。以孔子编纂的《春秋》为例，“有一事一辞者，亦有一事异辞者；又有一辞一事者，亦有一辞数事者……”(郭晓冬、曾亦，2017：97)——这也就是说，以“辞”为代表的价值取向往往是与特定的历史事件描述任务捆绑在一起的，事件—价值判断之间往往并不呈现出机器学习机制所乐见的“多对一”关系。也就是说，当我们真要将中国传统的历史记录文本当成初始资料“喂入”丘奇兰德式的神经元网络模型的话，那么，这样的模型是根本无法运作的，因为此类资料本身的“层次”已经高到无法对其进程逐级抽象的地步了(这个问题在后文中将被简称为“层次崩塌”问题)。

第二，如果神经计算模型的设计者要对儒家历史资料的原始内容进行“降低层次”的处理，以便清洗其中的价值判断的话，那么，又一个哲学义理问题又会迎面而来：仅仅从对于一个社会事件的纯粹物理描述中，我们是无法抽象出其价值属性的，除非我们已经获知了一个更大的评价语境(譬如，正是特定的语境知识，才使得我们赋予了“中途岛海战中美国海军飞行员对于日本航母的自杀式袭击”与“日军‘神风特工队’在战争后期对于美海军的自杀式袭击”以不同的意义，尽管从物理层面上看，两种行为的确是很相似的)。但这一点却会马上陷机器学习机制于两难：若这样的机制不允许其所处理的初始材料包含此类语境知识的话，那么，其进行抽象分类的结果就会与人类的常识大相径庭；若它允许初始材料包含此类语境知识的话，那么，此类语境知识自带的价值维度则会使得“层次崩塌”的麻烦重新出现。

第三，同样是为了避免“层次崩塌”的麻烦，一个丘奇兰式的机器学习机制就必须保证其输出是足够抽象且不带事实描述的。然而，通过类比历史掌故而对人物作出评价，却恰恰是儒家式人物评价的根本特点。譬如，陈寿在《三国志·吕布张邈臧洪传第七》中对东汉末年的“八厨”之一张邈(？—195 年)的总结性评价只有一句话：“昔汉光武谬于庞萌，近魏太祖亦弊于张邈”——换言之，陈寿试图经由“光武刘秀—魏祖曹操”以及“庞萌—张邈”之间的类比关系，含蓄地表达他对于张邈的负面看法。此段引文中的“谬”与“弊”虽带有明显的“高层”价值所向，它们却又同时附着于相关的“底层”史实之上——从技术上看，这就使得前面提到的“层次崩塌”问题重新浮现。这种局面无疑将再次使得“丘奇兰德们”陷入两难：他们要么就必须承认儒家“通过历史类比进行人物评述”的做法是无法被神经计算模型所模拟的，要么

就必须削足适履地修正儒家的表述习惯,以适应此类模型自身的技术特点。

上面的评估,无疑足以说明:任何一种试图用自然主义态度重建儒家德性论的技术路线,都不能通过一种"自下而上"的技术路径,而将儒家意义上的德性养成过程视为任何一种意义上的"模式识别"任务。毋宁说,对于任何一种典型的儒家式道德训诫样本来说,语义属性与价值属性都已经内在于其中,而无法被抽象掉了。因此,从哲学角度看,自然主义者必须接受语义属性与价值属性在原始材料中的"不可还原性",并在此基础上探求某种不预设描述层次之间的等级架构的新刻画方案。如何在这个新方向上进行探索,也便是下节所要触及的话题。

四、通过基于"儒家德性样板库"的隐喻性投射来获取德性

众所周知,隐喻是一种通过在表面上言及甲事而实际上由此涉及乙事的修辞手段。这一修辞手段对于儒家的历史叙述传统来说绝不陌生。现再从汉末历史中选取几例。汉末名臣第五种("第五"为复姓)被宦官势力陷害,后被江湖豪杰救走,遭到朝廷通缉。当时任徐州从事的臧旻(即前面提到的臧洪的父亲)上书天子为第五种辩护,并在相关文字里提到了"齐桓公宽恕曾用箭射过他的管仲"、"汉高祖宽恕曾为项羽效过力的季布"等历史故事,由此暗示当今天子也要对第五种"录其小善,除其大过"(相关资料收录于《后汉书·第五钟离宋寒列传第三十一》。为方便理解,笔者已对部分原文作了白话文改写)。不难看出,臧旻在其精心编排的修辞中,其实是通过对于齐桓公与汉高祖德性的提点而暗指时下天子的行为所应遵循的轨迹,尽管他没有明说天子就应当是当下贤君。很显然,这就是隐喻手法在儒家式政治规劝活动中的妙用。无独有偶,汉末名将孙坚在规劝司空张温斩杀桀骜不驯的董卓之时,也采用了类似的隐喻式修辞,即通过温习"穰苴斩庄贾、魏绛戮杨干"这样的故事来提示张温当下应做之事(《三国志·吴书·孙破虏讨逆传第一》)。至于一种更广泛意义上的隐喻机制,则在从汉代开始流行的谶纬系统中被全面地"体制化"了,譬如范晔在《后汉书·五行二》中对于"灾火"、"草妖"、"羽虫孽"、"羊祸"等自然现象的描述,实际上便包含了对于汉末衰微政局的一种"密集式"隐喻投射。

——那么,以上说的这些案例,又该如何被整合入德性伦理学的话语框架之中呢?

在前节中我们已经看到,丘奇兰德将德性伦理学"自然主义化"的要旨,便是通过神经计算模型来对道德刺激进行逐步抽象,并根据抽象的结果来对抽象进程进行反馈,由此使得系统获得正确的"抽象习惯"——此即"德性"。虽然我们已经知道了这种"逐层抽象"的技术思路是很难被运用到儒家德性训练的实际案例上去的,但至少就"通过特定训练样本形成某种具有规范性的推理习惯"这一大思路而言,我们依然可以在某个更恰当的技术平台上对其予以保留。依据笔者浅见,儒家人物评价模式对于隐喻式修辞的高度依赖,正好为构建上述这种"更恰当的技术平台"提供启发。譬如,我们可以按照这样的路线图来构建这种平台:

第一步:人类程序员通过史料阅读,手动建立一个"儒家德性样板语料库",而每一个语例都要按照如下格式标注各种参数的值:(甲)人名;(乙)典型事迹集;(丙)对每一典型事件背后当事人的道德决策进程进行心理重构;(丁)对每一典型事件进行整体上的道德价值词标注(有时一个复杂事件可以用几个价值词标注);(戊)对于该人物的总体德性评价。在整个"步骤一"中,对于环节(丙)中数据的采集可能是最为困难的,因为当事人的心理活动与道德决策过程往往很难在事后被复原。比较合理的处理方法是罗列出史料所记载的当事人面对特定任务时所需要满足的所有目标,然后根据他对于这些目标的实际取舍,反推出这些目标在其内部心理评价系统中的排位。而在环节(丁)中,我们将根据"击靶德性论"的精神,对每一事迹的成败给出价值评分,尔后再结合环节(丙)所给出的对于行为者意图的描述,构成某种综合评分(其综合标准是:"击靶"未成功的邪恶意图的综合道德评分会被拉高,而"击靶"未成功的善良意图的综合道德评分则会被降低,依此类推)。至于环节(戊)所涉及的对于人物德性的总体评价,则是对在环节(丁)中所出现的大量道德评注进行统计学抽象后的结果。此外,还需要读者注意的是,本步骤所涉及的"儒家德性样板语料库"反映的虽然是传统官方史书对于历史人物评价的一般性意见,但

我们并不试图假定此库中的所有参数设置会具有贯彻全库的逻辑自洽性(因为不同的儒家学者往往会对于同一人物自然会有不同的褒贬)。与这种宽容相对应,我们亦允许数据库营建方根据新资料对这个数据库进行修正与扩容。因此,这样的数据库便不会在任何意义上构成一个“公理系统”(笔者会在稍后提及实现这些技术理想的备选技术手段)。如图 3 所示。

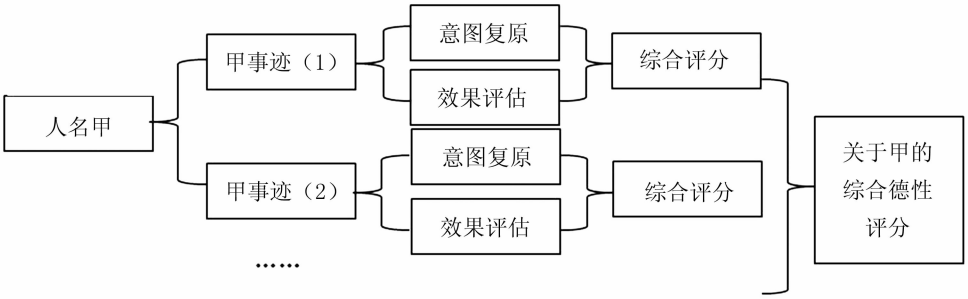

图 3　儒家德性样板库的建立(上)

第二步:系统必须对数据库信息进行自行整合,即在标注为“与当事人甲相关”的数据集与标注为“与当事人乙相关”的数据集的各个下属参数之间进行相似度计算。在理想情况下,一个已经具有强大类比推理性能的计算系统,将有能力自行在“齐桓公宽恕曾用箭射过他的管仲”与“汉高祖宽恕曾为项羽效过力的季布”这两个事例之间找到相关性,尽管这两个事例本来是属于两个不同的数据集的。由此,系统会自行形成与“宽恕”这种行为相关的典型语例集,由此构成对于以人名为核心词的语例集的二阶表征。如图 4 所示。

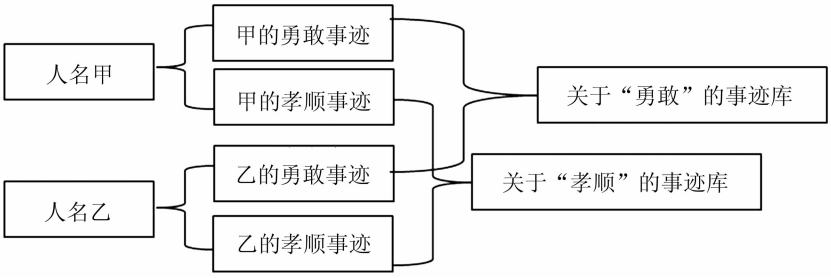

图 4　儒家德性样板库的建立(下)

第三步:向系统“喂入”一个新的虚拟道德情境,要求系统:(甲)通过类比思维方式,在前述“儒家德性样板语料库”的“一阶语例集”与“二阶语例集”中搜寻到特定的子集,并在这些子集与当下案例之间建立特定的隐喻投射关系;(乙)将语例库中的道德问题求解方案投射到当下案例中;(丙)给出问题的求解方案。

第四步:系统的“高层次评估模块”会对上述步骤所给出的求解方案进行评估,若其得分合格,则完成本轮训练,给出下一个道德案例进行深化训练;若得分不合格,则驱动系统重启“步骤三”,直到给出的解答符合评估要求(这里需要补充说明的是:通过头轮训练就使得系统输出满足要求的概率,恐怕是很低的。不难想见,如果样本库信息足够丰富的话,那么其所含的与当下情境貌似雷同的案例也就会非常之多,因此,系统就很可能会在初次选择备选的隐喻投射基准对象时“看走眼”)。

第五步:经过上述步骤而完成了大量道德案例训练的系统,其实已经具备了以恰当的方式将新的道德情境要素与道德样本库中的相关因素加以联系的能力,也就是建立恰当的道德隐喻投射关系的能力。我们可以认为:具备这种能力的系统已经具备了最初步的“德性”。

第六步:经过更长且更复杂的运行历史的此类系统,为了节省内部运作资源,将从自身处理特定问题的内部经验出发来面对新的道德情境,而不再大规模地求助于儒家道德样本语料库中的信息(而这一内部推理过程的简化之所以可行,也正是因为样本库中的德性样板已经通过复杂的学习历程而被系统

所内化了)。这样的系统也可以被视为是某种具有较为完整的德性(full-fledged virtue)的人工道德推理系统。

以上技术路线的实现,显然取决于对于合适的计算平台的选择。具体而言,这样的计算平台显然要具备对于自然语言的强大编码能力,以及对于类比思维的强大表征能力。同时,它还不能是任何一种意义上的基于公理集的封闭式推理系统,否则它就无法对应儒式道德推理的开放性与对于语境因素的敏感性。若说得再技术化一点,这样的计算平台应当能够像传统的亚里士多德式逻辑那样,支持某种基于"词项"的推理——因为对于类比—隐喻推理的表征任务而言,经由"本体与喻体对于某个间接词项的分享"来建立恰当的推理路径,其实是一条最经济的技术路线。而正如我们所看到的,神经计算模型是无法满足这些技术需求的,因为此类模型只能完成对于繁杂数据的识别任务,而无法在语义水平上直接进行逻辑推理(遑论相对复杂的类比推理与隐喻投射)。而基于公理系统的传统符号人工智能的进路,同样无法胜任我们在此所给出的任务,因为这样的技术进路在面对含糊、开放、易受语境因素影响的类比推理任务时,表现往往很拙劣——而更重要的是,此类技术进路对于弗雷格式现代逻辑(以及与之捆绑的"真值语义学")的依赖,使得其无法像词项逻辑推理系统那样规避所谓"框架问题"(徐英瑾,2011)①。而在笔者所知的范围内,目下在全球范围内,最可能将笔者所构想的"儒式德性训练模型"予以算法化的计算平台,乃是由美国天普大学(Temple University)的计算机科学家王培先生发明的"纳思系统"所提供的。纳思系统的英文全称为"Non-Axiomatic Reasoning System"(非公理推理系统),"NARS"为其缩写,"纳思"为该缩写的汉语音译②。大体而言,纳思系统乃是一个具有通用用途的计算机推理系统,而且在如下意义上与传统的推理系统有所分别:纳思系统能够对其过去的经验加以学习,并能够在资源约束的条件下对给定的问题进行实时解答。从技术角度看,纳思系统是由诸多层次的技术构建构成的,每个层次均有其自身的推理规则,而这些规则又都基于作为一种新词项逻辑的"纳思式逻辑"。整个系统之所以被说成是"非公理的",则是得缘于如下理由:尽管系统的构造者会在一开始为系统的每个层次预先设置一些推理规则,但他既不会将整个系统的知识库锁死,也不赋予知识库中的任何一个命题以公理的形式。毋宁说,纳思自身的知识库是可以随着系统的经验的增加而被不断修正和丰富的(这些修正本身则是在纳思推理规则的指导下进行的)。也正是在这个意义上,纳思的知识表征进路在实质上便不同于丘奇兰德所推崇的神经计算模型以及传统人工智能研究所推崇的符号规则进路,因为后二者均要求系统一开始就获得关于环境的充分知识(或接近于充分的知识)。由是观之,纳思进路颇有孔子所说的"君子不器"的品格,并天然与儒式推理方式相亲近(如果我们将"器"这个字重新解释为对于特定领域内的充分知识的执着态度的话)。当然,若我们真要着手经由纳思技术平台来构建本节所描述的儒式德性训练模型的话,由此所牵涉到的大量技术性讨论,恐怕是不能为这篇小文所包容的。有兴趣的读者可参看笔者经由纳思系统重构许慎"六书"构字理论的其它理论尝试(徐英瑾,2012),因为这些尝试所涉及的诸多技术细节,对于德性训练模型的营建来说也是通用的(同时,许慎的构字论本身,也可以被视为儒式隐喻式思维方式在文字学领域内的映现)。

五、余 论

正如笔者所反复提及的,本文重构儒家德性论的元哲学预设乃是"自然主义"的,即认定对于德性理论的重新表述不需要预设任何一种"超自然因素"(也就是无法被现代自然科学的话语框架所理解的因素)。从这个角度看,这种研究并不能被归类为任何意义上的"东西比较哲学"研究,因为笔者并不认为未经"自然主义"标准遴选过的西方哲学资源本身有资格成为评价、重述东方思想资源的"元语言框架",因为这些资源(如牟宗三的儒学重建所特别倚重的德国古典哲学资源)自身往往也没有完成被"祛魅化"

①这个问题的实质乃是"如何在外延化的真值语义学框架中表征自然词项之间的内涵关联性"。

②关于纳思系统的文献很多,其中最重要的是:Pei Wang. *Rigid Flexibility: The Logic of Intelligence*. Netherlands: Springer, 2006。

的进程。毋宁说，本研究对于计算机建模方式的引入，本身就是为了同时满足“使得被解释对象祛魅化”与“使得中国文化资源得以普世化”这两项目的，并由此切断将儒学重新神秘主义化与地方主义化的所有退路。依据笔者的管见，这种处理方式也可以使得儒家思想资源更好地接续马克思主义的唯物论立场，并使得由此获得的儒家新理论形态满足数码时代所提出的新理论期望，而不至于使得华夏民族古老的道德训诫沦为与后工业时代背景脱节的文化琥珀。

参考文献：

[1] 郭晓冬、曾　亦(2017). 春秋公羊学史. 上海：华东师范大学出版社.

[2] 徐英瑾(2011). 一个维特根斯坦主义者眼中的框架问题. 逻辑学研究，2.

[3] 徐英瑾(2012). 如何真正让电脑懂汉语——一种以许慎的“六书”理论为母型的汉语处理模型. 逻辑学研究，2.

[4] 杨泽波(2010). 从德福关系看儒家的人文特质. 中国社会科学，4.

[5] Robert Merrihew Adams(1999). *Finite and Infinite Goods*. New York: Oxford University Press.

[6] T. Chappell (2014). *Knowing What to Do*. Oxford: Oxford University Press.

[7] Paul Churchland(2007). *Neurophilosophy at Work*. Cambridge: Cambridge University Press.

[8] Chienkuo Mi, Michael Slote & Ernest Sosa(2015). *Moral and Intellectual Virtues in Western and Chinese Philosophy*. New York: Routledge.

[9] Edward Slingerland(2010). Toward an Empirically Responsible Ethics: Cognitive Science, Virtue Ethics, and Effortless Attention in Early Chinese Thought, in Brian Bruya (eds.). *Effortless Attention: A New Perspective in the Cognitive Science of Attention and Action*. Cambridge : The MIT Press.

[10] Michael Slote (2001). *Morals from Motives*. Oxford: Oxford University Press.

[11] Ernest Sosa (2015). Confucius on Knowledge. *Dao: A Journal of Comparative Philosophy*, 14(3).

[12] Christine Swanton(2003). *Virtue Ethics: A Pluralistic View*. Oxford: Oxford University Press.

[13] Pei Wang(2006). *Rigid Flexibility: The Logic of Intelligence*. Netherlands: Springer.

[14] Linda Zagzebski(2004). *Divine Motivation Theory*. New York: Cambridge University Press.

[15] Linda Zagzebski (2010). Exemplarist Virtue Theory. *Metaphilosophy*, 41(1/2).

Confucian Virtue Ethics, Neural Computation and Cognitive Metaphors

Xu Yingjin (Fudan University)

Abstract: Virtue ethics is an umbrella label for any positions in normative ethics according to which moral outputs of a moral agent is evaluated in terms of his dispositions of delivering such outputs. Confucianism can be perceived as an instantiation of contemporary virtue ethics for many reasons, especially due to the similarities between Confucianism and Linda Zagzebski's agent-based virtue ethics or Christine Swanton's target-based virtue ethics. And Confucianism can be further naturalized as an ethical position which can be fleshed out via some computable model on how an artificial moral agent should work. The first computable model evaluated in this paper is provided by Paul Churchland, who proposes to train a neural computational network to acquire "virtue". However, such proposal is believed to be defective since it cannot handle the problem of "the collapse of the hierarchy for levels of representations". A more preferable approach is going to be appealing to the metaphorical/analogical reasoning capacities based on theso-called "corpus on Confucian moral exemplars", and the technical realization of the whole approach is expected to be based on Pei Wang's "Non-axiomatic Reasoning System".

Key words: Confucianism virtue ethics; target-based virtue ethics; neural computation; metaphor

■**收稿日期**：2017-07-30

■**作者地址**：徐英瑾，复旦大学哲学学院；上海 200433。

■**基金项目**：国家社会科学基金一般项目(13BZX023)；国家社会科学基金重大项目(15ZDB020)

■**责任编辑**：何坤翁