中图分类号:TP18 文献标识码:A 文章编号:1004-8634(2017)06-0005-(08)
DOI:10.13852/J.CNKI.JSHNU.2017.06.001

具身性、认知语言学与人工智能伦理学

徐英瑾

(复旦大学 哲学学院,上海 200433)

摘 要: 在主流的人工智能伦理学研究中,很少有人意识到:将伦理规范的内容转化为机器编码的作业,在逻辑上必须预设一个好的语义学理论框架,而目前主流人工智能研究所采用的语义学基础理论却恰恰是成问题的。文章主张在人工智能与人工智能伦理学的研究中引入认知语言学的理论成果,并在此基础上特别强调"身体图式"在伦理编码进程中所起到的基础性作用。依据此番立论,并主张:机器伦理学的核心关涉将包括对人工智能的"身体"——而不仅仅是"心智"——的设计规范,即必须严肃考虑"怎样的外围设备才被允许与中央语义系统进行恒久的接驳"这一问题。

关键词: 人工智能伦理学;认知语言学;认知图式;具身性;语义学

随着人工智能技术的日益发展,对于此类技术产品的伦理学考量也逐渐进入学界的视野。概而言之,与人工智能相关的所有伦理学思考,都在国际学界被归入"人工智能伦理学"(ethics of Artificial Intelligence)的范畴,而这个学科分支本身又可分为两个小分支:"机器人伦理学"(roboethics)与"机器伦理学"(machine ethics)。前者的任务是对设计机器人的人类主体进行规范性约束,而后者的任务则是研究如何使得人类所设计的人工智能系统在行为上具有伦理性。这两个分支彼此之间既有分工上的分别,又有微妙的联系。两者之间的差别体现在:"机器人伦理"直接约束的是人类研究主体的行为,而"机器伦理"直接约束的是机器的行为。两者之间的联系又体现在:不通过"机器伦理学","机器人伦理学"的指导就无法落地;而没有"机器人伦理学"的指导,"机器伦理"的编程作业也会失去大方向。

不过,在当前人工智能伦理学研究中,很少有研究者意识到此类问题实质上乃是某种深刻的语言哲学-语言学问题的变种,而不能就事论事地在应用伦理学的层面上被谈论。而笔者的相关判断又是基于如下考量:如果我们要把用自然语言表达出来的伦理学规范——如著名的"阿西莫夫三定律"——转换为能为机器识别并执行的程序语言的话,我们就必须对人类的语言运作的本质有着一种预先的理论把握;而语言学家与语言哲学家对于人类语言机制的不同理解,则显然又会导致对于上述问题的不同解答方式。

此外,也正因为一般意义上的语言哲学-语言学问题在人工智能伦理学研究中的边缘地位,认知语言学关于"具身化"问题的见解也相应地被边缘化了。很少有人工智能伦理学方面的讨论触及

收稿日期:2017-09-01
基金项目:国家社科基金一般项目"自然语言的智能化处理与语言分析哲学研究"(13BZX023);国家社科基金重大项目"基于信息技术哲学的当代认识论研究"(15ZDB020)
作者简介:徐英瑾,上海人,教育部长江学者奖励计划青年学者,复旦大学哲学学院教授,博士生导师,主要从事认知科学哲学、人工智能哲学、知识论、维特根斯坦哲学等研究。

如下问题:伦理编程问题不仅仅牵涉软件的编制,而且还将牵涉"怎样的外围设备才被允许与中央语义系统进行恒久的接驳"这一问题。也就是说,依据笔者的浅见,机器伦理学的核心关涉将包括对人工智能体的"身体"——而不仅仅是"心智"——的设计规范。而为了支持这一看似"非主流"的观点,本文的讨论将始自于对如下问题的"务虚"式讨论:为何伦理学必须具有"具身性"?

一、从伦理学的"具身性"说起

"具身性"(embodiment)本是一个在认知哲学领域内使用的术语,其主要含义是指:人类认知的诸多特征都在诸多方面为人类的生物学意义上的"身体组织"所塑造,而不是某种与身体绝缘的笛卡尔式的精神实体的衍生物。如果我们将这样的观点沿用到伦理学领域之内,由此产生的"具身化伦理学"的核心观点便是:伦理学规范的内容,在相当大程度上便是为作为伦理主体的人类的肉体特征所塑造的。换言之,伦理学研究在相当程度上必须吸纳生物学研究的成果,而不能将自己视为与"肉体"绝缘的"纯精神领域"。

应当看到,将"具身性"与伦理学相互结合的观点,并不是西方伦理学研究的传统路数,甚至还与该领域内的思维定式相左。譬如,柏拉图就曾将"善"的理念视为超越于可感知的物理世界的最高理念,而康德则将道德律令视为某种凌驾于肉身领域的"绝对命令"。但随着演化论等自然科学思维范式逐渐进入伦理学领域,越来越多的具有自然主义倾向的伦理学家开始注意到了伦理学自身的生物性根基。正是基于此类考量,英国生态学家汉密尔顿(William Hamilton)在 1964 年提出了所谓的"亲属选择模型"。[1]根据该模型,在假定甲、乙两个生物学个体之间具有一定的遗传相似性的前提下,只要这种相似性与"乙从甲获得的好处"之间的乘积能够抵消"甲自身因帮助乙而遭到的损失",那么,使得互助行为可能的那些基因就会在种群中传播(这一规律,也在科学文献中被称为"汉密尔顿律")。或说得更通俗一点,依据汉密尔顿的理论,道德的生物学起源,很可能就是与"通过亲属的生存而完成家族基因的备份"这一隐蔽的生物学目的相关的。需要注意的是,汉密尔顿所给出的这种对于道德起源的描述看似抽象,其实已经触及"身体"对于伦理学的奠基意义。譬如,前述"汉密尔顿律"的起效,在逻辑上已经预设了一个生物学个体有能力将别的生物学个体识别为其亲属。而要做到这一点,辨认主体若不依赖于被辨认对象的身体形态的识别,则几乎是难以想象的。从这个角度看,道德意义上的"共情感"很可能便是以道德主体之间在身体方面的相似点为前提的。

对于上述的理论描述,有的读者或许会问:汉密尔顿的"亲属选择模型"又将如何解释人类对于非亲属的其他人所产生的同情感呢?实际上答案也非常简单:"基因的相似性"实质上是一个针对特定参照系才能够成立的概念。若以其他物种为参照系,整个人类都算是一个巨大的亲属组织,因此,你与地球上任何一个需要别人帮助的人之间都有着某种基因上的关联性。而按照"汉密尔顿律",只要这种关联度与"被帮助者从你这里获得的好处"的乘积能够大于"你因为帮助他而遭到的损失",那么利他主义行为就可以被激发。而在很多情况中,对于陌生人的很多帮助形式——譬如在网上向受灾群众捐献 10 元——所需要付出的生物学资源其实是微不足道的,这就使得"汉密尔顿律"所规定的相关条件在数学上变得容易被满足(换言之,"大于"左边的乘积实在太容易超过其右边的数值了)。或再换一个更通俗的说法:廉价的"助人为乐"行为的传播之所以并不是很难,就恰恰是因为这些行为自身所消耗的资源不多;而与此同时,人与人(尽管很可能彼此是陌生人)之间在身体层面上的起码的相似点却已经足以激发出微弱的"好感",以便催生那种微弱的利他性行为。与之相对应,代价不菲的利他主义行为却往往是建立在被帮助者与帮助者之间较密切的亲属关系之上的,并经由这种亲属关系所提供的更为强烈的"亲近感"驱动。

不过,笔者也承认,上述这种基于生物学考量的道德起源学说,并不能对人类所有的人际行为做出充分的描述,因为作为自然存在者与社会存在者的合体,人类的具体行为在受到生物学因素的制约外,还会受到社会-文化因素的制约与影响(譬如文化、生产方式、政治理念、宗教等因素对一个人的"亲密圈"的重塑效应)。但即使如此,生物学方面的考量依然会构成"文化重塑活动"的基本逻辑空间;换言之,文化重塑的方向本身必须首先是"生物学上可能的"。意识到这一点的美国哲学家麦金太尔便在《依赖性的理性动物》一书中,特

别强调了伦理学研究与生物学研究之间的连续性。他指出，如果我们将伦理学视为对人际关系根本规范的研究的话，那么，我们就无法忽略使得此类人际关系得以存在的下述基本的生物学前提：人类是一种离开了群体生活就必然会灭亡的物种，因为人类的身体具有一种生物学意义上的脆弱性。“我们是否能够存活，在相当程度上取决于别人（更别提繁衍了），因为我们经常遭遇如下困难：身体疾病或伤害、营养不足、精神疾病与困扰，以及来自于别人的入侵与无视……”[2](P1) 也就是说，按照麦金太尔的观点，人类道德规范中最为基本的那部分——如尊老爱幼、帮助弱小，等等——都是对于某些最基本的生物学需要的“再包装”，而不是脱离于人类的生物学实际的纯粹的“文化发明”。由此不难推出：如果在另外的一个可能世界中的人类的生物学习性与现有的人类不同（譬如，那个世界中的人类会像螳螂那样在交配之后吃掉“新郎”），那么，我们也就没有理由期望他们的道德规范内容与我们的道德规范基本一致了。

不难想见，如果这条“达尔文—汉密尔顿—威尔逊（E. O. Wilson，他的‘社会生物学’研究是汉密尔顿工作的全面升级化）[3]—麦金太尔”式的伦理学研究路数是正确的话，那么，此类思维方式就肯定会对人工智能伦理学产生直接的影响。这里需要被提出的最核心的问题便是：既然人工智能产品并不是任何一种意义上的“生物体”，我们又怎么保证此类产品能够经由其与人类身体的相似性而承载了人类所认可的道德规范呢？换言之，既然对吾辈而言人工智能体肯定是“非我族类”的，“其心必异”的结局难道不正是无法避免的吗？

不过，同样不容否认的是，至少对于主流的人工智能伦理学研究而言，人工智能制品因为其物理“身体”的不同而潜藏的对人类社会的伦理风险，并没有被充分注意到。譬如，著名的“阿西莫夫三定律”就表达了某种经过强制性的代码输入（而不是身体设计）以禁止机器人危害人类的企图。而在此路径的支持者看来，给相关的机器人配置怎样的“身体”反倒成为一个与机器伦理无涉的边缘性问题。此外，即使他们了解到从汉密尔顿到麦金太尔的整条“具身化的伦理学”的发展线索，恐怕他们也会以这样的一种轻描淡写的方式来打发“具身派”的见解：既然汉密尔顿所说的“利他主义基因”本身就是以自然选择的方式植入人类的一种强制性操作代码，那么，人工智能专家就完全可以自行扮演自然选择的角色，向机器直接植入这样的代码。他们或许还会补充说：既然自然选择本身并不是什么神秘的机制，那么，到底有什么自然选择能够做到的事情，我们人类做不到呢？

但在笔者看来，上面的辩驳是无力的。其一，自然选择机制的基本原理固然并不神秘，但是特定性状的演化历史的种种细节却很可能是难以被事后复原的，因为这牵涉相关基因与特定生态环境之间的复杂互动。因此，从非常抽象的角度看，如果将自然选择机制人格化为一个设计师的话，那么“他”对于伦理代码的编制路线便是“从下到上”（bottom-up）的，而阿西莫夫式的机器伦理代码的编制路线则是“自上而下的”（top-down），两条路径并不相似。其二，自然选择的过程不是一次完成的，而是通过“代码变异—引发显现型变化—参与生存竞争—筛选代码”这样的复杂流程，渐进式地积累各种遗传代码素材的。与之对比，阿西莫夫式的机器伦理代码设计流程，却试图通过某种一劳永逸的代码编制工作来杜绝未来可能发生的一切伦理风险，这无疑就需要设计者具备像上帝那样的预见力。而我们都知道，人类是永远无法扮演上帝的角色的。其三，自然选择的过程所积累的核心信息虽然是以基因代码的方式被加以保存的，但是在具体的生存竞争中，这些代码必须外显为身体的性状才能够兑现其生存价值。也就是说，至少对于生物体而言，其遗传代码本身就具有一种针对“具身性”的明确指向，而这种指向在阿西莫夫式的伦理编码里是找不到的。其四，也是最重要的，自然演化的“设计产品”是我们能够看到的（我们自己就是这样的“产品”）；而根据“阿西莫夫三定律”所设计出来的成熟的人工智能产品，我们却还没有看到。更有甚者，根据瓦拉赫（Wendell Wallach）与艾伦（Colin Allen）的分析，[4](P95) 在日常语境中对于“阿西莫夫三定律”的执行将不可避免地导致矛盾。譬如在面对一个人正在残害其他人的场面时，以下这两条法则就很难被同时执行：“机器人不得在与人类的接触中伤害人类”；“机器人不得在目睹人类被伤害时袖手不管”。瓦拉赫与艾伦就此指出：只要我们允许机器人以比较大的自主性来独立应付种种困难的伦理局面，我们就无法指望其行为能够同时满足“阿西莫夫三定律”的僵化规定；相反，如果我们硬是

要将这些僵化规定以代码的形式植入机器的话，我们就不能指望它们的行为输出时是富有智能的。很显然，这是一个两难困境。

从瓦拉赫与艾伦的分析中，我们还可以从语言哲学与语言学的层面发现自然选择所进行的“伦理编码”与阿西莫夫式的“伦理编码”之间的又一重要差异。不难看出，自然选择所遴选出来的基因编码组合本身是不带语义的，而兑现这些基因组合的生物体的身体表现同样是不带语义的——赋予其语义的，乃是人类观察者对于这些表现的事后描述。因此，对于自然选择来说，就不存在着“如何将带有语义内容的伦理规范分解为具体算法”这样的问题；而与之相对比，阿西莫夫式的机器伦理编制者却不得不面临这样的难题(因为“阿西莫夫三定律”本身无疑是带有语义的)。因此，除非机器伦理学家们将自身的程序编制工作奠基在一个恰当的语义学理论之上，否则，此类工作就无法解决瓦拉赫与艾伦所指出的那类“灵活性与原则性不可兼得”的困难。

在本节的讨论中我们已经得到了两方面的结论：首先，我们已经看到，伦理规范自身很难摆脱“具身性”的规制；其次，我们发现阿西莫夫式的机器伦理编制工作既没有意识到“具身性”的重要性，同时也缺乏一个使得其自身的语义内容落地的语义学基础理论。而要将这里所说的具身性考量与语义学考量结合起来，我们就需要一个合适的理论媒介。这一媒介就是认知语言学。

二、认知语言学的“具身性”对于人工智能伦理学的启示

这里笔者之所以要提到认知语言学，乃是因为它提供了一个将前面所提到的“具身性原则”与语言学理论相互结合的范例。认知语言学的代表人物之一雷考夫(George Lakoff)[5](P9)就曾概括过认知语言学的意义观与传统语义学的意义观之间的不同。笔者按照自己的理解将其列为表 1 概括于下：

认知语言学提出了很多技术概念，以便将表 1 中所提出的意义观加以具体化。一个非常具有代表性的技术概念是“认知图式”(cognitive schema)。“图式”一词的古希腊词源“σχήμα”有“形状”的意思，而在认知语言学的语境中，“图式”指的则是“一系列语例中的共通性在得到强化后所获得的一些抽象的模板”。[6](P23)在认知语言学中，此类“抽象模板”往往按照“意象式”(imagistic)的方式来加以把握；而所谓的“意象式”结构，本身乃是“前概念”的，是具有一定的“可视性”的。譬如，英语“ENTER”(进入)这个概念就可以被分析为数个意象图式的组合，包括“物体”(object)、“源点—路径—目标”(source-path-goal)与“容器—容纳物”(container-content)。三者结合的情况如图 1 所示。

表 1　认知语言学的意义观与传统语义学的意义观对比表

传统语义学的意义观	认知语言学的意义观
存在着判定何为正确的世界结构的“上帝之眼”	没有这样的“上帝之眼”，不同的语用共同体会给出不同的世界描述
所有人都使用相同的概念结构	任何概念结构都是依赖于特定视角而存在，并因为视角的不同而彼此不同
意义理论的核心范畴是“真”与“指称”，而此两者又关涉符号与世界中对象的关系	意义理论的核心关涉乃是对特定对象的“范畴化”(categorization)，即使得某类内部对象“从属于”另一类内部对象的心智运作机制。外部世界中的外部对象并非意义理论所关心的
心智与身体相互分离	心智与身体不可分
情绪没有语义内容	情绪有语义内容
语法是纯形式的	语法是语义内容的衍生物
推理是先验与超越一切具体领域的	推理敏感于被涉及的语义对象的具体语义内容

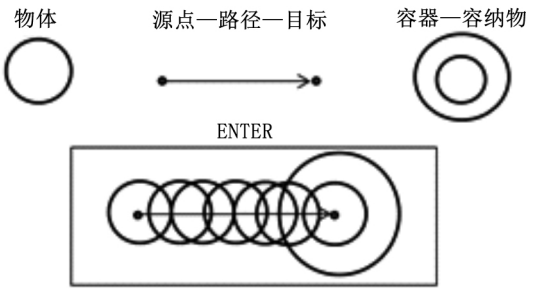

图 1　关于“ENTER”的认知图式形成过程的图示[6](P33)

这样的“可视化”图式当然包含了明确的身体指涉。说得更清楚一点，这样的概念图示预设了概念的使用者具有这样的身体经验：自主移动身体，从一个源点出发，沿着一定的路径，进入一个“容器”，并由此使得自己成为一个“容纳物”。换言之，一个从来没有移动过自己的身体，甚至从来没有观察到其他物体之移动的语言处理系统，恐怕是无法真正把握“ENTER”图式，并由此真正

把握“进入”这个概念的含义的。

而在从属于“图式”的各个“图式要素”中,“辖域”(scope)这个术语亦与“具身性”有着密切的关联。在认知语言学语境中,“辖域”指的是一个目标概念在被聚焦时,语用主体的注意力在语义网中所能够覆盖的周遭范围。其中与核心区域较近的周遭范围叫“直接辖域”,而注意力的最大范围边界便是“最大辖域”。譬如,对于“肘部”这个概念来说,其直接辖域就是“胳膊”,其最大辖域就是“身体”;而对于“铰链”这个概念来说,“门”就是其直接辖域,“房屋”则是其最大辖域。同时,也正因为任何一个被聚焦的对象与其直接辖域之间的连带关系,我们的自然语言表达式往往只允许该对象与相关直接辖域的概念名彼此连缀为复合名词(如“肩胛骨”“门铰链”等),而不允许该对象跳过直接辖域,与更宽泛辖域的概念名彼此连缀为复合名词(如“体胛骨”“房铰链”等)。[6](P64~65)

而之所以说基于“辖域”的认知语言学叙述方式体现了“具身性”的要素,乃是因为:任何“辖域”的存在均有赖于其边界的存在,任何“辖域”边界的存在又有赖于认知主体视野范围的大小,而认知主体视野范围的大小则最终又取决于其身体的特性。换言之,“辖域”的特征归根结底还是由认知主体的身体特性所塑造。为了更形象地说明这一点,我们不妨设想一下《格列佛游记》中“大人国”与“小人国”居民可能的概念认知图式所具有的“辖域”。譬如,对于一个大人国的居民来说,他所注意到的整个人类房屋恐怕就只有正常人类所看到的一块豆腐那么大,在这样的情况下,“铰链”对于“门”的直接从属关系将因为显得过于“微观”而变得可被忽略。而与之相对比,对于小人国的居民来说,正常人类尺度上的“房屋”却可能因为“过于宏大”而无法成为其所聚焦的某些(对人类而言的)微观对象的辖域;甚而言之,小人国的认知语言学家们恐怕还会开发出一系列对正常人类而言匪夷所思的概念,以便对他们眼中围观对象的辖域进行描述,如在将“铰链”作为聚焦对象的前提下,提到“铰链近侧”“铰链远侧”,等等。

那么,以上说的这些,与伦理学有什么关系?与人工智能伦理学又有何关系?

首先可以肯定的是,伦理学所研究的社会规范本身往往就带有“身体图式”的印记。让我们不妨来想想在周遭社会中遇到的种种社会规范所具有的语言表达吧!比如:“这是军事禁地!禁止入内!”(这个表达预设了关于“进入”的身体图式)“行车时不能挤占公交车道!”(这个表达预设了“挤占”这个概念的身体图式)“不许占据别人的财物!”(这个表达式预设了“占据”是一个将远离身体的非辖域转化为其近侧辖域的动态过程)“不许杀人!”(这个表达式预设了被涉及的人类身体的确具有终止别的人类身体的生物学机能的物理能力)不难想象,如果上述这些关于身体图式的预设全部被抽空的话,那么我们就很可能会凭空造出一些让人不知所云的社会规范,如:“永远不能挤占银河系之外的空间!”“永远都不能通过吐口水的方式来淹死长颈鹿!”……这也就从认知语言学的角度印证了汉密尔顿—麦金太尔路线的伦理学研究思路,即:伦理学规范的内容是某种生物学需要的或直接或间接的再包装,而不是脱离了这些需要的纯精神臆造物。

而上述的研究思路一旦被推广到人工智能伦理学——尤其是机器伦理学——领域内,就会立即触发如下的问题:既然没有任何科学方面的理由使得“小人国”或“大人国”尺度上的智能机器人无法被制造出来(且不论这么做在伦理上是否合适),那么,我们又如何保证这样的机器人所具有的“认知图式”会与人类的“认知图式”彼此合拍呢?而如果这种“合拍性”无法被保证的话,我们又如何保证同样建立在机器人自身身体图式之上的机器人的伦理规范能够与人类既有的伦理规范相合拍呢?而如果后一种“合拍性”也无法被保证的话,我们又如何保证这样的智能机器人不会对人类既有的社会秩序构成威胁呢?

为了使笔者所表达的上述疑虑不显得那么空洞,在此还想表达两个补充性意见。其一,在笔者看来,任何智能机器人——如果其具有真正意义上的全面智能的话——所具有的语言智能,都应当包括对于身体图式的识解能力(无论它的身体构造是怎样的,也无论它是如何获取这种识解能力的)。之所以如此判断,则是基于如下推理:机器人所使用的符号若要与外部环境产生有效的、富有灵活性的互动的话,就必须对使用特定符号的典型语境有所把握,而身体图式恰恰是浓缩了此类典型语境信息的最佳推理中介。因此,认知语言学家关于人类认知图式的很多理论,至少就其哲学精神而言是适用于未来人工智能体的。至于如何找到合适的编程手段来体现认知语言学的原则,则是另外一个话题了。其二,在笔者看来,

如果两类智能体(无论是人造的还是自然的)在身体图式上存在着时空尺度上的巨大差异的话,那么这两个话语体系之间的转译成本就会显得非常巨大,甚至有时会变得不可转译(我们不妨设想一下:倘若蚂蚁也会说汉语的话,它们"爬"概念的身体图式就会与我们人类的"爬"有很大的差异)。这一观察对于伦理规范编制的直接影响就是:机器人很可能会因为无法识解人类伦理规范所隐藏的身体图式,而无法识解整条规范(比如,被完全做成海豚状的水下机器人会因为无法理解"踩踏"概念而无法理解"禁止踩踏人类"这条规范的意义)。退一步讲,如果这些机器人语言智能的发达程度已经达到了允许其通过内部的类比推理来间接把握人类身体图式的地步,那么,它们由此所理解的人类社会规范对其而言也只具有一种抽象的意义(而非实践的意义),因为这样的规范实在离它们自己的"生活形式"太远。在此情况下,我们就很难指望这样的智能机器人会严肃地对待人类的社会规范,并在此基础上成为人类所期望的工作与生活中有用的帮手。

基于以上的讨论,在笔者看来,为了防止种种对人类不利的情况出现,机器伦理学家就必须预先阻止"完整意义上的语义智能"与"与人类的时空尺度迥然不同的身体构造"这两项因素在同一个机器人身上的结合。而要做到这一点,从逻辑上看,我们就只有三个选项:

选项一:姑且可以去建造与人类的时空尺度迥异的机器人(比如非常微小的纳米机器人),但是不赋予其高级语义智能,即不赋予其在复杂环境下独立、灵活地做出决策的能力(在这种情况下,此类机器人的智能显然是不足以丰富到足以将"身体图式"予以内部表征的地步的)。

选项二:在将人类意义上的灵活智能赋予机器人的时候,必须要保证其身体界面与人类的身体界面没有时空尺度与性能表现上的重大差异。或说得更清楚一点,这样的机器人不能比人类跑得快太多或强悍太多,但也不能比人类慢太多或脆弱太多。甚至我们要鼓励更多的人形机器人的开发,使得机器人与人类之间能够形成基于"身体上的彼此承认"的"共情感"。套用麦金太尔式的"需要伦理学"的话语框架,也可以这么说:我们必须在硬件构造上就使机器人产生对其他机器人特别是人类的"需要",就像人类社会中的任何一个成员在生物学意义上需要他人才能够生存一样。

选项三:我们可以将富有灵活智能的机器人与比较"愚笨"的机器人临时组合起来,让前者去操控后者,就像人类自己也会临时地去操控相对缺乏自主智能的机械一样(在这种情况下,两类机器人之间组合的"临时性",可以依然保证高智能机器人自身认知图式的"拟人性"不被破坏)。但需要注意的是,切不可将这两类机器人的临时组合长久化以催生某种对人类不利的新的认知图式,就像人类自己也只能在特殊情况下才允许士兵去使用真枪实弹一样。

当然,除了以上三个选项之外,笔者也不排除:在某种特殊情况下(比如在某些对人类而言极度危险的作业环境下),我们将不得不把"完整意义上的语义智能"与"与人类的时空尺度迥然不同的身体构造"这两项因素予以永久性地结合。但即使如此,我们也至少要保证此类机器人与主流人类社会没有广泛的空间接触,以维护人类社会的安全。

如果用一句话来概括笔者论点,那便是:太聪明的人工智能并不构成对人类的威胁。且毋宁说,太聪明的人工智能与超强的外围硬件设备的恒久组合形式,才会构成对人类的威胁,因为与人类迥异的身体图式本身就会塑造出一个与人类不同的语义网络,并由此使人类的传统道德规范很难附着于其上。基于此推理,人工智能伦理学的研究方向应当"由软转硬",即从对软件编制规范的探讨,转向研究"怎样的外围硬件才允许与人工智能的中央处理器进行接驳"这一崭新的问题。

不过,正如笔者已在前文所提及的,笔者的上述观点是与人工智能伦理学的主流意见相左的。在下一节中,我将回过头来再对这些主流见解进行简要的评述。

三、主流人工智能伦理学研究对于具身性的忽视

首先应当看到的是,就当下的发展状态而言,"人工智能伦理学"依然是一门非常不成熟的学科分支。实际上,即使在世界范围内,推动"人工智能伦理学"研究的并不是学院内部的力量,而主要是各国官方与企业的力量,其背后的动机与企图也带有非常强的应景性,并不是立足于学科发展的内部逻辑。譬如,有军方背景的人工智能伦理学家主要关心的是"能够自动开火的机器人"所应当遵循的伦理规范问题;[7](P49~67) 欧洲议会在

2016年发布的一份建议性文件甚至讨论了将在欧盟范围内把被普遍承认的民权准则赋予机器人的问题。[8]在笔者看来,这两项问题的提出均已超越了目前人工智能的实际发展水平,并带有明显的"消费议题"的嫌疑,因为在认知语言学的相关学术洞见还没有被人工智能的编程作业消化的前提下,现有的人工智能系统的语义表征能力实际上都是不足以编码任何人类意义上的道德规范的——无论这样的人工智能系统的使用环境是军用的还是民用的。更有甚者,在夸张当下人工智能发展水平的前提下,近年来物理学家霍金在各种场合都在散布"人工智能威胁论",[9](P62~104)并在公众中制造了一些不必要的恐慌。在笔者看来,这种"忧患意识"就好比是在一个核裂变的物理学方程式还未被搞清楚的时代就去担心核战的危险,的确只是现代版的"杞人忧天"罢了。

另外,也正是因为参与上述讨论的各界人士其实并没有将有关人工智能研究的相关语义学与语言哲学问题想透,他们忽略了"身体图式"对于伦理规则表征的奠基性意义,并因为这种忽略而错过了人工智能伦理学研究的真正重点。譬如,研究军用机器人的相关伦理学专家所执着的核心问题——是否要赋予军用机器人以自主开火权——本身便是一个不着边际的问题。在笔者看来,只要投入战争的机器人具有全面的语义智能(这具体体现在:它能够理解从友军作战平台上传送而来的所有指令与情报的语义,能够将从其传感器中得到的底层数据转化为语义信息,并具有在混杂情报环境中灵活决策的能力,等等),在原则上我们就可以凭借它们的这种语义智能对其进行"道德教化",并指望它们像真正的人类战士那样知道应当在何种情况下开火。在军事伦理的语境中更需要被提出的问题乃是"我们是否允许将特定的武器与机器人战士的'身体'直接、恒久地接驳",因为这种直接接驳肯定会改变机器人战士的身体图式,并由此使人类对它们的"教化"变得困难。

而在相对学院化的圈子里,至少在人工智能伦理学的范围内,对于具身化问题的讨论其实也不是很够。即使是哲学家德瑞福斯(Hubert Dreyfus)多年来所一直鼓吹的"具身性的人工智能"路径,[10]在具体技术路线上也与认知语言学其实并无多大交集,而且他也尚未将关于"具身性"的观点延展到人工智能伦理学的领域。至于前面已引用过的瓦拉赫与艾伦合写的《道德机器——如何教会机器人"对"与"错"》一书,[4]虽然在行文中的确时常流露出"要使机器人的道德决策机制更接近人类"的思想倾向,却也并没有在评述业界既有技术路线的同时,给出一条富有独创性的技术路线来实现这样的思想倾向。因此,无论在学院外部还是内部,本文所提出的以"身体图式构建"为理论基石的人工智能伦理学研究路径,的确算是一条比较新颖的思路。不过,限于篇幅,本文并没有勾勒出将此类身体图式构建与具体的计算机编程作业相互结合的技术路线图。而相关的研究,显然需要另外一篇论文去完成了。

参考文献:

[1] Hamilton, W D.The GeneticalEvolution of Social Behavior [J]. Journal of Theoretical Biology, 1964,7(1).

[2] MacIntyre, Alasdair. Dependent Rational Animals: Why Human Beings Need the Virtues[M]. Open Court, 1999.

[3] Wilson, E O. Sociobiology: The New Synthesis[M].Cambridge, MA: Belknap Press, 2000.

[4] Wallach, W and Allen, C. Moral Machines: Teaching Robots Right from Wrong[M]. Oxford: Oxford University Press, 2009.

[5] Lakoff, G. Women, Fire, and Dangerous Things: What Categories Reveals about the Mind [M]. Chicago: University of Chicago Press,1987.

[6] Langacker, Ronald. Cognitive Grammar: A Basic Introduction[M]. Oxford: Oxford University Press, 2008.

[7] Lin, P, et al.. Robots in War: Issues of Risks and Ethics [A]. Capurro, R. and Nagenborg, M. Ethics and Robotics [C]. Heidelberg: Akademische Verlagesgesellschaft AKA GmbH, 2009.

[8] European Commission. European Parliament, Committee on Legal Affairs. Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics[Z/OL].http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN.

[9] Dreyfus, H. Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian[A]. Julian Kiverstein and Michael Wheeler. Heidegger and Cognitive Science[C]. New York: Palgrave Macmillan, 2012.

[10] Matyszczyk, Chris. Hawking: AI Could be "the Worst Thing Ever for Humanity"[Z/OL]. https://www.cnet.com/news/hawking-ai-could-be-the-worst-thing-ever-for-humanity/.

(下转第57页)

㉜同上书,第 98 页。
㉝陈松长主编:《岳麓书院藏秦简(五)》。
㉞张家山二四七号汉墓竹简整理小组编:《张家山汉墓竹简(释文修订本)》,第 99、104、105 页。
㉟何有祖:《里耶秦简牍缀合(六则)》,来源:http://www.bsm.org.cn/show_article.php?id=1765,2012 年 12 月 24 日。
㊱(汉)司马迁:《史记》卷八十七《李斯列传》,中华书局 1959 年版,第 2561 页。
㊲(晋)陈寿:《三国志·魏书》卷三《明帝纪》,中华书局 1959 年版,第 107 页。
㊳(晋)陈寿:《三国志·魏书》卷十二《司马芝列传》,中华书局 1959 年版,第 388 页。
㊴陈伟主编:《里耶秦简牍校释(第一卷)》,第 70、195 页。
㊵陈松长主编:《岳麓书院藏秦简(五)》。
㊶(汉)班固:《汉书》卷一下《高帝纪下》,第 63 页。

A Complementary Study on the Submitted Reports on Trial of Qin Dynasty ——From the Perspective of the Yuelu Qin Bamboo Slips

WEN Junping
(Yuelu Academy, Hunan University, Changsha 410082, China)

Abstract: The submitted reports on trial of Qin dynasty mainly included two parts, one was the doubtful cases, and the other was the specific cases and cases in which the suspects held special status. Based on the ordinances of the Yuelu Qin bamboo slips, "Dang", the proposal for judgement, had its regular format and it was usually attached on "Ju". "Dang" was usually found in the criminal cases reports submitted to Tingwei, the main judicial officer, from the lower prefecture. And, due to the specific feature of the cases, the prefecture officials had no right to make the final decision, and Tingwei or the emperor were eligible to make the final judgement.

Key words: submitted report on trial, dang, submitted criminal cases of Qin dynasty, Yuelu Qin bamboo slips

(责任编辑:知　鱼)

(上接第 11 页)

Embodiment, Cognitive Linguistics and Ethics of Artificial Intelligence

XU Yingjin
(School of Philosophy, Fudan University, Shanghai 200433, China)

Abstract: The issue on how to make ethical codes "computable" in order to build artificial moral agents (AMA) is fundamentally a semantic issue on how to semantically represent the relevant norms by appealing to some proper algorithms. However, unfortunately, the mainstream semantic theories employed in artificial intelligence (AI) are problematic in this aspect or another; and more unfortunately, the mainstream studies in the ethics of artificial intelligence haven't realized the relevance of semantic considerations to ethics yet. What the author of this paper intends to propose include three points. Firstly, cognitive linguistics (CL), especially its notion of "cognitive schema", could be a useful resource for building the needed semantic framework for AI; secondly, due to the fundamental status of the notion of "embodiment" in the whole CL-narrative, we can hardly imagine any semantic representations of ethical codes in the case of of designing AMAs if these representations are not based on certain physical features of the "bodies" of AMAs; thirdly, due to the preceding considerations, issues like "what kind of peripheral equipment is allowed to be permanently connected to the information-processing center of a certain AMA" should be put on the table for all ethicists of AI.

Key words: ethics of artificial intelligence, cognitive linguistics, cognitive schema, embodiment, semantics

(责任编辑:知　鱼)