武漢大學學報(哲学社会科学版)
WUHAN UNIVERSITY JOURNAL (Philosophy & Social Science)
第73卷第6期 2020年11月
Vol. 73 No. 6 Nov. 2020 005~014
DOI:10.14086/j.cnki.wujss.2020.06.001

通用人工智能需要在私人语言的层面上进行知识表征吗
——来自大森庄藏的启发

徐英瑾

摘 要 通用人工智能语境中的私人语言,指的是这样一个意思:表征A在系统甲那里的知识表征方式与同一个表征在系统乙那里的表征方式必然会有所差异。因此,在预设推论主义语义学自身有效性的前提下,A在甲中的意义集,总会有一个子集(无论这一子集有多小)仅仅为甲自身所拥有,而无法被任何一个别的系统所拥有。因此,任何一个与甲不同的别的系统,都无法彻底地理解甲对于A的意义把握方式。很显然,这样一种将机器表征与哲学史上的"第一人称哲学"传统相互结合的思路,是无法见容于后期维特根斯坦对于私人语言的著名反驳的。而为了与后期维特根斯坦论战,日本哲学家大森庄藏的思想资源便具有了很高的引用价值,因为他本身的哲学就可以被视为"维特根斯坦的话语方式与胡塞尔的思想内核"的日本式混合体。在对大森的哲学进行面向机器表征问题的重建的过程中,对于局域性原则与历史性原则的引入也是题中应有之义,以便为通用人工智能语境中建设私人语言的必要性提供辩护。而非公理性推理系统(纳思系统)所提供的技术手段,则会为这种想法的技术落地提供可能。

关键词 大森庄藏;通用人工智能;私人语言;非公理化推演系统
中图分类号 B222;B84 **文献标识码** A **文章编号** 1672-7320(2020)06-0005-10
基金项目 国家社会科学基金重大项目(15ZDB020);教育部哲学社会科学研究重大课题攻关项目(19JZD010)

传统的符号主义进路的人工智能研究在哲学预设上素常带有客观主义意蕴,意即相关编程作业所设计的虚拟模型,被假设为对于客观世界的某个特定面相的摹写(这条路线,下面简称为"第一个版本的客观主义")。实际上即使在基于大数据的深度学习模型中,这样的一个预设也没有真正退场,因为深度学习所模拟的,无非就是大量的人类个体在特定输入与输出之间的映射习惯,并由此在某种康德式的意义上依然属于客观主义路线(康德主义意义上的客观性所意指的即大量主观性的交错重叠之处。此路线,我们下面简称为"第二个版本的客观主义")。从语言哲学的立场上看,第一个版本的客观主义属于柏拉图—弗雷格路线的意义观,而第二个版本的客观主义则与后期维特根斯坦的意义观有一些关联。然而,如果我们将智能的标准提高到创造性与个性的标准,上述两种客观主义的意义观都无法解释为何某些个体的智能能够超越客观性的窠臼而另辟蹊径。这就倒逼我们去恢复某种版本的主观主义的意义观传统,并借此机缘,对诸如后期维特根斯坦哲学这样的客观主义意义理论进行反思。在这个问题上,日本分析哲学家大森庄藏复活第一人称哲学传统的努力,其实是颇有参考价值的。

一、导论:如何在通用人工智能研究的视角中恢复“第一人称哲学”的尊严

西方认识论的叙述视角,素有“基于第一人称的视角”与“基于第三人称的视角”之分。前一路线的代表人物有普罗塔哥拉、笛卡尔、洛克、胡塞尔等,后一条路线的代表人物则有柏拉图、黑格尔、赖尔、后期维特根斯坦等。按照一般人的观点,人工智能(AI)的研究,应当与后一条路线更有关联。相关的论据如下:AI 所需要的编程语言与界面语言,都需要有足够的清晰性,并尽量消除可能的歧义——而满足这一要求的语言,将不得不基于所谓的“第三人称的视角”,因为只有该视角才能容纳主体之间的相互检查与相互沟通,由此消除个体观察视角带来的偶然性因素,最终使得语言表征变得足够清晰明白。

但只要我们结合 AI 发展的具体实践,就会发现,上述观点应当只适用于所谓的专家系统,而不是时下如火如荼的联结主义—深度学习技术路径,遑论还在雏形中的通用人工智能研究。

先来看专家系统(expert system)。所谓专家系统,就是“一个以特定方式编制的计算机程序,以使得其能够在专家的知识层面上运作”[1](P63-64)。具体而言,典型的专家系统的研制方法,是先将一个特定知识领域内的专家知识用逻辑语言加以整编,然后利用逻辑推理规则推演出对用户有用的特定结论。很明显,此类系统所涉及的专家知识本身,往往便是那些经过特定领域的人类学科共同体的反复锤炼而被普遍认可的知识,因此当然是基于第三人称视角的。

但联结主义—深度学习的技术路径就不是这样了。该技术的实质便是用数学建模的办法建造出一个简易的人工神经元网络结构。一个典型的此类结构一般包括三层: 输入单元层、中间单元层(在深度学习框架中,这样的中间单元层可以包含大量亚层,数量从 4 个亚层到上百个亚层不等),以及输出单元层(参看图 1)。输入单元层从外界获得信息之后,根据每个单元内置的汇聚算法与激发函数,决定是否要向中间单元层发送进一步的数据信息。中间单元层再将信息加以处理,输送给输出层,输出层再将输出结果与人类给出的标准答案比对,根据比对结果决定是否要启动反向传播算法来调整神经网各单元之间的信息传播路径的权重。这样的系统在如下三重意义上是不支持基于第三人称视角的知识表征的:(1)在此类技术路径中,对于完整的语言表征的处理,已然被分解为大量的亚符号运算,而不像专家系统那样,一开始就将特定的命题知识固化为系统的知识库的内容。(2)又恰恰因为在联结主义-深度学习的系统中,并没有命题性表征的线性传递路线,故此,就连此类系统的构建师自身,亦缺乏对于特定信息在系统内部的处理路径的追踪能力。毋宁说,他们只能通过瞎蒙的方式来调整系统的参数,以图使得系统达到令用户满意的信息处理水准。而此类系统的这种黑箱性质无疑使得在第三人称视角中对于它们的运作机理的可解释性成为一个大难题。(3)此类技术路径所需要的训练数据往往需要人工标注,以便产生用以判断系统所输出的识别标签是否正确的标准答案——而此类标注又往往会固化特定人类标注员的偏见,由此形成整个系统的算法偏见,并最终进一步破坏某种更具普遍意义上的基于第三人称视角的知识表征。

然而,以上说的这些并不意味着联结主义—深度学习的技术路径能够成为前述“基于第一人称视角”的知识论路线的自觉的工程学承载者,因为对于第三人称视角中的明晰性的排除,未必就必然意味着自动获取那种具有第一人称视角中的明晰性(如笛卡儿主义者所说的“我思”所呈现出的那种明晰性)。毋宁说,在这种技术路径中,由于一个关于自我的心理学建模的匮乏,此类系统其实是缺乏一种真正意义上的第一人称视角的。其具体工程学表现是:在这样的系统完成训练后,这样的系统既缺乏对于自身组织结构的元知识的表征能力,也缺乏对于这样的结构的自我修正能力,而只能胜任在某类特定输入与特定输出之间的映射建立任务。

那么,以上说的这些是否意味着“基于第一人称的视角”的哲学认识论路线,就在原则上与 AI 无缘呢? 答案是否定的。实际上,如果我们讨论的 AI 具有通用人工智能(Artificial General Intelligence,简称 AGI)的特征的话(也就是说,这样的系统应当能够胜任各种任务,而并非仅仅只能执行特定的任

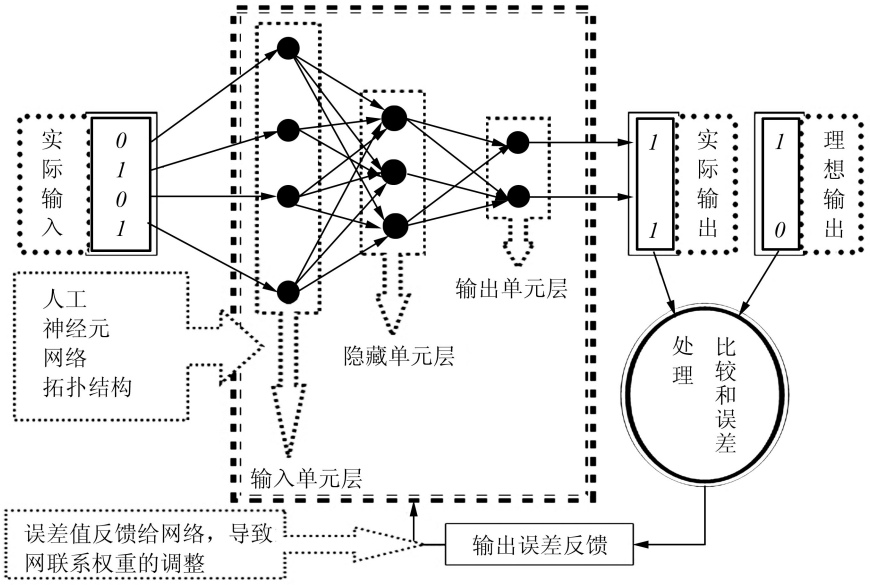

图 1　一个被高度简化的人工神经元网络结构模型

资料来源: 笔者自制

务),那么,在 AI 语境中对于上述认识论路线的兑现,至少是具有明确的工程学价值的。其道理是:如果我们指望此类 AGI 系统能够在开放的环境下进行自主化运作的话(譬如希望此类系统能够在火星等恶劣环境下,在独立于人类遥控的前提下自主处理各种突发状况),那么,这样的系统就需要有能力随时根据最新的情况更新自身的知识库,并对未来还未发生的新情况进行合理的预期。这同时也就意味着:这样的系统是应当具有记忆、怀疑、展望等典型人类心理能力的等价物的,并因此具有某种起码的主观性。进而言之,由于不同的 AGI 系统所各自面临的生存环境的差异,基于不同环境互动历史的生存策略就会在不同系统的"主观性"面相上打下自己的烙印,由此使得第一人称视角成为 AGI 系统的某种不可或缺的特征。

然而,在 AGI 语境中对于第一人称视角的尊重与相关建模活动,无疑会遭遇到一个非常明显的哲学反驳,即这种尊重无法见容于后期维特根斯坦对于私人语言的批判。说得具体一点,如果私人语言被定义为"一种指涉仅仅为言说者自己所知(而无法为他人所理解)的东西(特别是言说者的直接的私人感觉)"的话[2](P135),那么,在 AGI 语境中对于第一人称视角的重建,似乎也等于给出了这样一种承诺:对于两个特定的 AGI 系统 A 与 B 来说,存在着某些表征能够被 A 更为充分地理解,却不能被 B 同样充分地理解——反之亦然。但既然后期维特根斯坦是明确反对私人语言的可能性的,那么,看来他也不可能认为在 AGI 语境中对于第一人称视角的重建是有希望的。

很明显,唯一能让我们摆脱此困境的办法,便是去论证在这个问题上维特根斯坦可能是错的。为了增加此类论证的力度,本文将引入日本战后最重要的分析哲学家大森庄藏(1921-1997)的相关思想资源。而之所以引入大森哲学,则是基于如下考虑:(1)大森明确反对维特根斯坦的反私人语言论证,因此便是本文立论的天然盟友;(2)作为日本最早系统研究维特根斯坦的学者之一,大森本人反对维特根斯坦的话术结构本身就是继承自维氏哲学的,因此,基于大森哲学的反驳路线就会具有更强的说服力;(3)大森的哲学还包含一个系统化的说明,以便解释如何从具有第一人称视角特征的表征出发,营建出具有第三人称视角特征的表征系统。因此,他对于维特根斯坦立场的反驳,并不会让他自己的哲学成为一种唯我论——相反,他完全有能力对维特根斯坦所重视的公共语言的起源进行一种大森式的说

明;(4)大森的相关思想是有机会在 AGI 的技术语境中得到大致的模拟的——通过这种模拟,我们也便能初步勾勒一种具有第一人称特色的机器表征的大致样貌。

下面,笔者就将逐步展开上述论点。

二、大森是如何利用维特根斯坦去反对维特根斯坦的

大森庄藏虽然是日本著名的维特根斯坦专家,但他在 1971 年发表的著作《语言、知觉与世界》中,却明确表达了他对于基于第一人称视角的认识论道路的同情态度——而这种态度,显然是与后期维氏对于私人语言的敌对态度唱反调。大森写道:

> 为了确认他口中的"红色的印象"与我的印象是否是相同,我们必须将他的印象与我的印象相互比较。为了进行这种比较,我将不得不接受他的印象。但是因为我实际上并没有通向他人的感知的路径,这种比较是难以实现的。为了经验到他人的知觉,我就必须变成他;但由于施加于我自身之上的种种限制,这一点是无法被实现的。而且,这个问题,在原则上就是无解的。即使我是"暹罗连体人"之一也枉然:考虑到我就是我自己,而不是其他任何人,并且我也不能变成我的连体人兄弟,故此,我依旧无法感知到我的连体人兄弟所感知到的。[3](P13-14)

大森庄藏在上文中所提到暹罗连体人案例,显然是参照了维特根斯坦在《哲学研究》中用到的同一案例。维氏原文如下:

> 只要"我的疼痛同他的疼痛一样"这话是有意义的,那么,两个人也就可能有一样的疼痛(我们甚至可以想象两个人在相同的——不仅是相应的——部位感到疼痛。例如暹罗连体人就是这样)。[2](P98)

由此看来,尽管大森与维氏都利用了暹罗连体人的案例,二人的深层用意却是南辕北辙的。在大森那里,此案例是为第一人称视角的基本性提供注脚的,而在维氏那里,它却是为第三人称视角(或曰"公共视角")的基本性提供辩护的。大森本人对于他与维氏的上述区别,自然是心知肚明的。他是以如下方式为他自己利用"暹罗连体人"案例的方式提供辩护的:

> 如果我没有弄错的话,在前文中所阐述的观点,可以被认为是维特根斯坦的观点。然而,我无法接受上述观点所蕴含的如下观点:个人的心理体验必须被公共化。[3](P17 注 1)

对于上述论述,大森进一步补充论述如下:

> 无论多少信息可以经由语言而从外部环境而取得,且无论语言本身多少次得到了调整,所有的这一切针对语言的学习与调整,毕竟都是基于某人的具体目的的。对于我而言,语言的意义只有从我的视角出发才能得到理解。甚至所谓他人的语言,也无非就是我所能理解的语言。譬如说,当别人说什么"红色的大轿车"的时候,无论他是如何理解"红色的"这个词的,而且,无论他本人的对应感觉究竟为何,我本人对于"红色的"的理解,却总是基于我对于该字眼的理解之上的,而且,如何指派此词的意义,也总是取决于我。纵然语言可以被众人所分享,并因为它被众人所分享而成其为语言,理解一种语言却总是某人自己的事情。[3](P21)

现在,我们就从 AGI 的角度,来重构大森论证。这样的论证有两个。第一个论证是基于不同的信息处理系统的空间局域性的,而第二个论证则是基于不同的信息处理系统的运行历史的特异性的(不过,其中第二个论证的有效性,在一定程度上是有赖于第一个论证的)。论证一可分为以下六步:

1. 对于大量语词——如红色——的理解,都脱离不了具体的感性样本,如一辆红色的大轿车。这一点对 AGI 系统也不例外,因为与特定感性样本脱离的符号输入,会在 AI 语境中造成所谓的语义奠基问题(grounding problem)[①]。

2. 任何一个 AGI 的信息处理系统,都需要针对其所处的特定物理环境的外部特征给出特定的反应。因此,这样的系统都具有物理意义上的局域性。

3. 因为 2,一个机器人的物理位置的局域性,就决定了其传感器所捕捉的信息具有特定性(譬如,一个处在此处的 AGI 机器人所捕捉到的关于"红色大轿车"的视觉信息,就会在色调、亮度等维度方面与另一个处在彼处的 AGI 机器人所捕捉的同类信息有所不同)。

4. 由 2 与 3,我们可得知: 当两个不同的 AGI 机器人都试图掌握同一个符号——如红色——的含义时,其获得的用以训练的基础数据,肯定是彼此不同的(尽管这种差异可能也是很细微的)。

5. 由此我们就可导出: 对于不同的 AGI 系统 A 与 B 来说,他们各自基础输入数据集之间的差异就会导致它们所要把握的概念的含义的区别,而无论这种区别有多细微。

6. 所以,对于 A 来说,其所理解的红色就总会与 B 所理解的红色有所差异,而无论这种差异有多细微。

论证二可分为以下四步:

1. 任何 AGI 系统的运行历史,都积累了其与外部环境互动时所产生的经验,因此,关于此类历史的数据,乃是此类系统在开放环境中进行决策的重要参考。

2. 由于 1,任何一个 AGI 系统对于任何一个概念的把握方式,都会参考其运行历史中对于此概念或者相关概念的理解方式(如果这种历史数据的确存在且可以被调取的话)。

3. 由于论证一的第二步所提到的局域性原则,任何两个不同的 AGI 系统的各自运行历史参数都会彼此不同。

4. 由于 3 与 2,对于某个公共符号甲来说,任何一个 AGI 系统对于它的把握方式,都会与另外一个 AGI 系统对于它的把握方式有所不同。

需要指出的是,上述两个论证的结论虽然都殊途同归,但它们都不支持这样一种观点——任意两个 AGI 系统之间都不能完成有效的沟通——因为 A 与 B 之间的有效的沟通不意味着"A 与 B 之间能够彻底地相互理解"。事情毋宁说是这样的: 对于符号甲的理解方式来说,A 与 B 各自的理解方式只要彼此重叠到一定程度,就能够进行比较有效的沟通了,而不论它们各自的理解方式在重叠区之外还有哪些分殊。大森还用了一个视觉隐喻色彩浓郁的术语,来描述这种使得公共交流得以可能的机制: 叠加描绘(日语"重ね描き")。具体而言,在日语中,"重ね描き"的意思就是先在画纸上铺上基础色,然后再在此基础上逐层加色——而透过画师所加上的每一层新色,人们依然可以看到下面的旧色。通过这个隐喻,大森实际上想讨论的乃是作为基础语言的第一人称视角语言与作为附着色的第三人称视角语言之间的关系。大森本人曾用看杯子为例,来具体说明这一点。众所周知,我们在看杯子的时候,不同的人从不同的角度所看到的杯子,都是杯子的不同侧显样式罢了,而每一个这样的侧显样式又都带有林林总总的第一人称视角色彩。与之对比,众人所谈论的那个作为物理对象的杯子,却是分明带有第三人称视角的色彩的。那么,我们是如何从对于杯子的特殊性侧显出发,进抵那个作为物理对象的一般性的杯子呢? 大森的答案就是: 诸多的关于杯子的侧显样式彼此重合叠加,然后我们的心智机器又各自将物理意义上的杯子构造为基于上述元素的一个理想化集合。然而,在这一集合中,每一个参与构造的特殊的杯子侧显却没有被淹没,而是依然可以像透出的底色一样,隐隐约约地显示出自己的本来面目[3](P91)。也就是说,在大森看来,透过公共语言所产生的各种约定,我们可以看到每个具体的言说者自己的个性化的语

① 这个问题可以被通俗地理解为"认知系统中的符号如何获取其意义"这样一个问题。

言把握方式,这一点也不妨碍公共语言在一个更高层面上的运作。

下面笔者将从 AGI 的角度,谈谈如何在技术背景中实现大森的想法。

三、纳思系统中的私人语言

笔者在 AGI 背景中对于大森想法的技术重构,将援引国际 AGI 活动的代表之一、华裔计算机科学家王培先生发明的非公理推演系统(简称为纳思系统)。换言之,在本节中,笔者将向读者呈现通过纳思系统重构 AGI 意义上的私人语言的可能性。考虑到任何语言——包括私人语言——的表达都是以判断为起点的,而最简单的判断无疑是主—谓判断,所以,我们的讨论也将始于纳思系统对于主—谓判断的表征方式。

与一阶谓词逻辑对于基本判断的表征方式不同,从纳思系统的立场上看,一个判断之中的主—谓差别,并非是自足的专名与未被满足的命题函项之间的差别,因为在纳思系统的基本术语表中,像“命题函项”这样的带有明显的弗雷格色彩的概念是没有地位的。毋宁说,纳思系统所运用的逻辑——纳思逻辑——与亚里士多德式的词项逻辑之间的亲缘关系,要明显强于其与弗雷格式的现代逻辑之间的亲缘关系。在纳思系统中,一个最简单的判断或信念乃是由两个概念节点构成的,比如,乌鸦(RAVEN)和鸟(BIRD)这两个节点。在纳思系统的最基本层面 Narese-0 上,这两个概念节点经由继承关系加以联接,该关系本身则被记作“→”。这里的继承关系可以通过以下两个属性而得到完整的定义:自返性(reflexivity)和传递性(transitivity)。举例来说,命题“RAVEN → RAVEN”是永真的(这就体现了继承关系的自返性);如若“RAVEN → BIRD”和“BIRD → ANIMAL”是真的,则“RAVEN → ANIMAL”也是真的(这就体现了继承关系的传递性)。这里需要注意的是,在继承关系中作为谓项出现的词项,就是作为主项出现的词项的内涵集中的成员(因此,在上述判断中,“鸟”就是“乌鸦”的内涵的一部分),而在同样的关系中作为主项出现的词项,就是作为谓项出现的词项的外延集中的成员(因此,在上述判断中,“乌鸦”就是“鸟”的外延的一部分)。换言之,与传统逻辑哲学家的思虑不同,在纳思的推理逻辑中,内涵并不代表某种与外延具有不同本体论地位的神秘的柏拉图式对象,而仅仅是因为自己在推理网络中地位的不同而与外延有所分别。

大量的此类纳思式主—谓判断,则由于彼此分享了一些相同的词项而构成了纳思语义网,如图 2 所示:

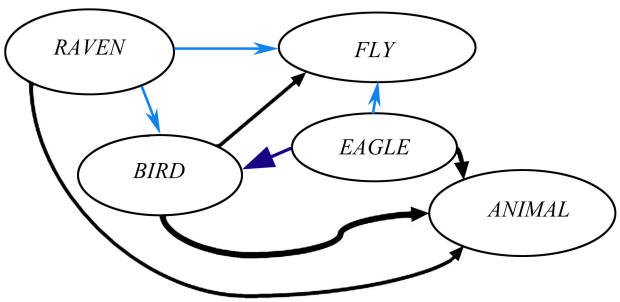

图 2 一个简易的纳思语义网

现在我们就来看看,上面的技术路径是如何使得私人语言得以在 AGI 的背景中得到刻画的。前文已经提到,在 AGI 的背景中说一种机器表征系统具有私人语言的表达,是说这样的一个机器表征系统中至少有一个这样的子集:该子集只能被一个特定的 AGI 系统所充分理解,而不能被任何一个其它的 AGI 系统所充分理解。而在这里,所谓的“充分理解”的定义则是这样的:某个 AGI 系统甲能够充分理

解另外一个 AGI 系统所给出的表达式 A，当且仅当：A 在甲的内部表征中所呈现出的推理结构所具有的拓扑关系，与 A 在乙的内部表征中所呈现出的推理结构所具有的拓扑关系完全重合。譬如，图 2 体现了系统甲对于概念 BIRD 的理解方式的话，那么，只有当另一个系统乙对于 BIRD 的理解方式能够完全不差地体现为图 2 的样子的时候，我们才能说乙能够完全理解甲对于 BIRD 的理解方式。否则，对于乙来说，系统甲对于概念 BIRD 的理解方式就带有私人语言的色彩。反之亦然。

但笔者将立即指出，恰恰是因为任意两个 AGI 系统之间的概念推理结构几乎不可能完全一致，故而，私人语言在 AGI 的内部表征中的出现，便是某种常态。之所以说任意两个 AGI 系统之间的概念推理结构几乎不可能完全一致，其基本理由便在于上节已经提到的所谓的局域性原则与历史性原则（这两个原则分别对应于前述论证一与论证二的基本前提）。这两个原则本身，完全可以在纳思系统中得到复演。

先来看局域性原则。在 AGI 语境中，该原则说的就是：每个 AGI 系统都有自己特定的空间处所，并因为这种差异而造成其传感器所获取的外部信息之间的差异。这种差异将进一步造成系统对于相关概念的理解方式的差异。需要注意的是，在纳思系统中，我们可以把一个前符号层面上的心理学意象（image）——在 AGI 的语境中，意象可以姑且通过一个像素矩阵而得到表示——视为一个词项，只要它能够被用以谓述其它词项，或被其它词项所谓述。比如，在（a）和（b）的例子中，不同的关于乌鸦的意象（在这里它们也都扮演词项的角色）就构成了词项“乌鸦”的外延，因为它们都可以被“乌鸦”所谓述。至于（a）和（b）本身，则构成了两个特殊的纳思系统语句（顺便说一句，作为这两个语句各自的主项，乌鸦图像的头指向方向彼此相反，以暗示二者在感性层面上有些许差异）：

图 3　不同感性图像与同一符号联结样式之一

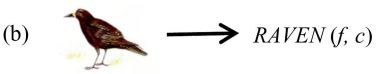

图 4　不同感性图像与同一符号联结样式之二

而这两个语句本身甚至还可以被融入整张纳思系统语义网之中，以构成图 5。

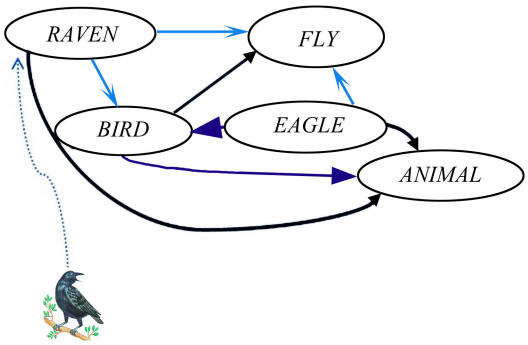

图 5　一张包含“心像”的纳思语义网

很明显，由于前述局域性原则，两个 AGI 系统所获得的针对特定概念的外延式感性示例，就往往会

产生彼此的差异,无论这种差异有多么的细微。这种差异,又会导致一个 AGI 系统所把握的一个特定概念的意义,不同于另一个系统所把握的同一个概念的意义。但这又是为何呢? 这是因为根据纳思系统的语义学,任何一个纳思概念的意义都是由其内涵集与外延集所构成的——而在当下的语境中,针对概念 RAVEN 的外延集显然就是指两个不同的 AGI 系统的传感器所获得的关于乌鸦的不同的图像模式。换言之,这二者在感知层上的差异,倒过来就造成了对于系统甲而言的 RAVEN 的意义与对于系统乙而言的 RAVEN 的意义之间的差别,由此进一步导致两个系统对于同一个概念的理解方式之间的区别。在这样的情况下,若系统甲在使用 RAVEN,且系统乙也观察到甲在使用 RAVEN,那么系统乙很可能就会激活自身使用同样一个概念时所依赖的局域语义网,并期待系统甲也可能按照同样的推理路径,来得到与自己的推理相同的推理结果。但又恰恰因为甲与乙各自所使用的 RAVEN 概念之间的确有着一定的语义学差异,甲对乙的上述期待往往会在某些情况下落空,而意识到这一点的系统甲便会由此发现其与对方的信息沟通产生了某种程度的挫折。从这个角度看,对于 AGI 系统的内部表征来说,私人语言的私人性不仅是可能的,而且甚至是可以被系统自身的高阶表征能力所自觉表征到的。

讨论完了局域性原则对于机器表征的私人性的影响,我们再从历史性原则出发来进行讨论,以便达到同样的论证效果。根据该原则,任何一个 AGI 系统都有自己特定的信息处理历程——因此,系统甲对于概念 A 的理解方式,自然会受到这样的特定历史信息的影响,由此产生与系统乙对于 A 的理解方式的偏差。说得更具体一点,信息处理进程所施加的此类影响,将主要从两个方面产生:

第一,由于不同的环境交互历史,不同的 AGI 系统会进入不同的时空坐标,由此反复激活局域性原则,导致其各自所获得的特定概念的外延集的彼此差异。这种差异,当然会导致不同的 AGI 系统所把握的同一个概念的把握方式之间的差异。

第二,AGI 系统所面临的外部环境,不仅包括物理环境,而且也包括信息环境。譬如,不同的 AGI 系统会通过与网络信息的接触,来得知一个特定概念的不同高层归属方式。举例来说,由于某些机缘,系统甲会被告知“病毒”属于广义的“生物”,而由于另外一种机缘,系统乙又会被告知“病毒”并非“生物”。由此,在不同的 AGI 系统那里,特定概念的内涵集也会产生彼此的差异。这就无疑会加剧不同的 AGI 系统对于同一个概念的把握方式之间的差异。

读到这里,针对笔者的上述旨在强化机器内部表征之私人性立论,敏锐的读者或许会提出这样两种反驳:

反驳一:假设我们故意将两个 AGI 系统的外部物理与信息环境都调整到完全彼此一样,这是不是就能够使得它们彼此的充分理解成为可能呢?

反驳二:反过来说,如果两个 AGI 系统彼此彻底理解,会因为同一个概念在二者那里的推理结构的差异而变得不可能的话,我们又有何理由说系统甲所说的 RAVEN 与系统乙所说的 RAVEN 是同一个概念呢? 若这一保证无法给出的话,我们又如何防止从否定“系统之间的彻底可理解性”出发,得出“系统之间不能进行任何沟通”这一摧毁性的结论呢?

先来看怎么对付反驳一。笔者的见解是:假设我们不仅将两个 AGI 系统的外部物理与信息环境都调整到完全一样,甚至还额外地将这两个系统的内部参数与先天知识也都调整到完全一样的话,那么,我们也就没有理由说这两个系统的确就是两个系统了。它们其实就是一个系统。而莱布尼茨的不可分辨原则,则将为上述的判断提供哲学依据。根据该原则,两个对象的所有属性若都完全一致,则我们就没有理由说这是两个对象,而只能说这是一个对象。但需要指出的是,此类的前提条件本身几乎是无法被满足的:一是只要两个 AGI 系统各自的传感器被放置到不同的空间坐标内,它们就会得到关于同一个概念的不同外延性示例(而且,只要两个 AGI 系统的确是彼此差别的,一般而言,二者就很难完全分享共同的物理环境信息);二是即使我们仿照前文所提到的暹罗连体婴儿的思路,让两个系统分享同样

的传感器，要再进一步将二者的软性信息环境调整到彼此彻底相同的地步，不仅在实践上是困难的，而且从社会需求的角度上看，也会造成毫无意义的资源浪费。

再来看反驳二。笔者回应这一反驳的思路，将基于大森哲学提出的叠加描绘概念。换言之，系统甲与系统乙对于概念 A 的各自理解方式固然不同，但只要它们发现以下两个条件能被满足，那么它们就能确定：它们的确是在谈论同一个 A：

条件一：甲所谈论的 A 在记号层面上与乙所谈论的 A 是一致的；

条件二：甲基于 A 所做的语义推理路径，在足够大的程度上与乙基于 A 所做的语义推理路径有所重合，以便为一种基于各种私人语言的叠加描绘奠定基础。

当然，关于如何进一步限定条件二所涉及的“在足够大的程度上”这一修饰语，我们还需要引入更多的实践层面上的规定。由于这些规定很可能是属于语用学的研究领域并因此涉及诸多的语用细节，限于篇幅，笔者就暂时不将此类问题展开了。需要注意的是，纳思系统本身已经提供了足够丰富的推理规则，以判断两个概念之间的相似程度是否足够高[4]（P100）——这就为我们满足条件二所提出的要求提供了算法基础。因此，在 AGI 的语境中实现大森关于叠加描绘的哲学设想，其实是颇有希望的。

四、总结性评论

本文立论的基本预设是：AGI 系统在处理任何任务时，处理资源（特别是信息资源与时间资源）之不足，乃是某种常态。因此，真正的 AGI 系统需要正视这种不足，并像具体的、个别的人类一样，时刻面对着特殊的物理环境与信息环境，而不能狂妄地认为自己能够一劳永逸地应对所有的物理环境与信息环境。所以，它们也要像具体的、个别的人类一样，具有自己的个性，甚至具有类似于人类心理活动的内部信息处理模式，以便以更为灵活的方式处理内部的知识表征。需要注意的是，这样的预设是不为主流的 AI 研究无论是专家系统还是深度学习所分享的，因为主流的 AI 技术路径都预设了系统的正常运作所需要的基本信息（无论是公理化的专家知识，还是带有标注的大量训练数据）乃是充分的（或是接近充分的）。在这样的预设下，人类意义上的心理活动便成为某种冗余了：因为对于某种全知者来说，它是不需要回忆那些过去的事情的（因为过往与当下一样，都无差别地摆在了它的眼前），它也不需要期待那些即将发生的事情（因为它已经确知了哪些事情即将发生）。同理，它也不需要因为对于自己的知识匮乏的自觉而感到恐惧。然而，不幸的是，主流 AI 技术的上述预设却肯定是错误的，因为此类路径所依赖的公理化知识也好，带有标注的训练数据集也罢，毕竟都是来自人类的，而人类自身却并非是全知者。换言之，主流 AI 技术对于全知者地位的僭越，本身就包含着对于外部环境自身的易变性与复杂性的无知，而这种无知又往往会在外部环境与系统本身的技术秉性发生冲突时，让系统出丑。与之相比，我们希冀中的 AGI 系统，则因为包含了对于人类心理活动的模拟，而使得系统自身能够在外部环境发生变化时，自主修正自己的知识图谱。此外，又恰恰因为这种模拟将在一个自觉的层面上凸显系统的知识表征的视角主义特征（根据视角主义，根本就没有独立于任何经验者之独特视角的中立性经验），所以，任何一种独立于特殊视角的知识表征，也无法见容于这样的 AGI 研究思路了。私人语言的私人性，只不过就是上述思路的一项题中应有之义罢了。本文的另一个目的是想提醒读者：当像 AGI 这样的工程学研究试图从哲学获取思想启发时，未必一定要按照“知名度差的哲学家不如知名度大的哲学家”的遴选原则。譬如，本文立论所参考的大森哲学，在国际范围内的知名度远不如其所试图反驳的后期维特根斯坦哲学。但大森哲学基于观察者视角的知识建构思路，与 AGI 的研究思路更为暗合，反而可能对 AGI 更有参考价值。同时，“大森哲学的知名度不高”这一事实，恐怕也与其主要作品缺乏外语译文有关（本文引用的大森思想材料，都直接采自日语原本），而与其自身的思想价值的高低没有关系。身为以汉语为母语的研究者，我们更当摆脱对于西语文献的过度依赖与崇拜心理，而应当关注同样处在“汉字文化圈”内的日本哲学工作者的努力。除了大森哲学，笔者窃以为，至少就日本哲学对于 AGI 研究的启发意

义而言,九鬼周造的“偶然性”哲学与西田几多郎的“场所逻辑”都是极有挖掘价值的。不过,对于它们的阐发与利用,显然不是本文的任务①。

参考文献

[1] A. Edward, Feigenbaum, Pamela McCorduck. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. Boston: Addison-Wesley, 1983.
[2] 维特根斯坦. 哲学研究. 陈嘉映译. 北京: 商务印书馆,2016.
[3] 大森庄藏. 言語 知覚 世界. 东京: 岩波书店,1971.
[4] Pei Wang. *Rigid Flexibility: The Logic of Intelligence*. Netherlands: Springer, 2006.

Is It Necessary to Construct Private Language in the Knowledge Representation of an Artificial General Intelligence System?

Some Remarks Inspired by Ōmori Shōzō

Xu Yingjin (Fudan University)

Abstract In the context of Artificial General Intelligence (AGI), the existence of private language implies that how a certain symbol is presented in systems' inner data-bases is different from one specific AGI system to another. Accordingly, when the validity of the "inferentialist semantics" is assumed, for any symbol S and two different systems A and B, the set of meaning-encoding units concerning S in A has to embrace a subset which is only a part of A's data-base but not of B's. Hence, for any system other than A, it cannot fully understand how S is construed in A's case. The revival of the idea of private language in AGI is definitely conflicting with later Wittgenstein's criticism of the possibility of private language. For refuting Wittgenstein, Ōmori Shōzō's insights may be fairly helpful here, given that his own philosophy can be viewed as a hybrid of Wittgensteinian terminology and a Husserlian core. In the process of reconstructing Ōmori's arguments, the "principle of locality" and the "principle of historicity" will also be introduced to strengthen the necessity of bringing the idea of private language into AGI. As to how to technically realize the very idea in AGI, Non-Axiomatic Reasoning System (NARS) may provide a promising approach.

Key words Ōmori Shōzō; Artificial General Intelligence (AGI); private language; Non-Axiomatic Reasoning System (NARS)

■ **收稿日期** 2020-08-08
■ **作者简介** 徐英瑾,哲学博士,复旦大学哲学学院教授、博士生导师,北京大学哲学系外国哲学研究所兼职研究员;上海 200433。
■ **责任编辑** 何坤翁

① 对于九鬼哲学与人工智能之间关系的阐述,可以参看徐英瑾在《山西师大学报(社会科学版)》2020 年第 9 期发表的《苏、日、欧人工智能发展错误决策后的哲学迷思》一文。对于西田哲学与一般意义的认知科学之间关系的阐述,可以参看徐英瑾在《学术月刊》2015 年第 8 期发表的《西田几多郎的“场所逻辑”及其政治意蕴——一种基于认知语言学的解读》一文。